REPORT DOCUMENTATION PAGE

AFRL-SR-AR-TR-03-

0249

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT T
		1 Feb 00 - 31 Jan 03

4. TITLE AND SUBTITLE
Estimation, Control, and Redundancy Management for Uncertain Networks of Cooperating Agents

5. FUNDING NUMBERS
F49620-00-1-0154

6. AUTHOR(S)
Jason Speyer

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Californi, LA
Mechancial & Aerospace Eng.
38-137 Engr IV, Box 951597
Los Angeles, CA 90095-1597

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
4015 Wilson Blvd, Room 713
Arlington, VA 22203-1954

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
F49620-00-1-0154

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED

20030731 034

13. ABSTRACT *(Maximum 200 words)*

The coordination of spatially distributed systems of cooperating agents, which perform an assigned mission in the presence of uncertainty and system faults, is an important emerging technology. The actions and health of these distributed systems depend upon the information that can be communicated and the knowledge of the current capabilities of all cooperating agents. Methodologies for the distribution of estimation and redundancy management functions over the dynamic network of cooperating agents were developed, leading to effective team strategies Progress has been made on various aspects of the distributed systems problem. From the fundamental level we investigated the decentralized control problem with constrained communication. In parallel the allocation of transmit power in wireless networks was a focus of study into the decentralized control problem because it has a simple structure and the information communicated is constrained. In the area of health monitoring new robust analytical redundancy methods have been developed which detects, identifies, and reconstructs sensor, actuator and plant faults. A robust multiple-fault filter is developed based on a performance measure from which the desired detection subspaces are approximately constructed. This detection filter formulation, which includes uncertainty, is the bases for single-fault time-varying, decentralized detection filters, and fault magnitude reconstruction. An innovative application of distributed detection filters methodology is to the target track association problem. Finally, the distributed estimation problem was addressed by considering elements of the relative navigation problem among distributed vehicles.

14. SUBJECT TERMS			15. NUMBER OF PAGES
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

FINAL REPORT

ESTIMATION, CONTROL, AND REDUNDANCY MANAGEMENT FOR UNCERTAIN NETWORKS OF COOPERATING AGENTS

Grant No. F49620-00-1-0154

Jason L. Speyer
Department of Mechanical and Aerospace Engineering
UCLA

Abstract

The coordination of spatially distributed systems of cooperating agents, which perform an assigned mission in the presence of uncertainty and system faults, is an important emerging technology. The actions and health of these distributed systems depend upon the information that can be communicated and the knowledge of the current capabilities of all cooperating agents. Methodologies for the distribution of estimation and redundancy management functions over the dynamic network of cooperating agents were developed, leading to effective team strategies

Progress has been made on various aspects of the distributed systems problem. From the fundamental level we investigated the decentralized control problem with constrained communication. In parallel the allocation of transmit power in wireless networks was a focus of study into the decentralized control problem because it has a simple structure and the information communicated is constrained. In the area of health monitoring new robust analytical redundancy methods have been developed which detects, identifies, and reconstructs sensor, actuator and plant faults. A robust multiple-fault filter is developed based on a performance measure from which the desired detection subspaces are approximately constructed. This detection filter formulation, which includes uncertainty, is the bases for single-fault time-varying, decentralized detection filters, and fault magnitude reconstruction. An innovative application of distributed detection filters methodology is to the target track association problem. Finally, the distributed estimation problem was addressed by considering elements of the relative navigation problem among distributed vehicles. Exact statistical solutions to the pseudorange equations in GPS and an efficient nonlinear filter based on multiple hypothesis sequential probability ratio tests for resolving the integer ambiguity in differential carrier GPS were developed and extended.

Accomplishments

The following accomplishments in the study of cooperative agents are divided into three categories; decentralized control, robust fault detection filters for distributed analytic redundancy management systems, and nonlinear estimation applied to relative GPS navigation among moving vehicles.

1. Decentralized Control

1.1 A Stochastic Decentralized Control Problem with Noisy Information

A simple decentralized stochastic control problem is considered where the non-classical nature of the information pattern is induced by the uncertainty of the information transmission in the system [1, Appendix A]. This is in fact a reformulation of the Witsenhausen counter-example, where the first station is allowed to send its' information to the second station through a noisy channel. Non-convexity of the problem in this new formulation has been established and it is shown how this formulation relates to a classical problem and the Witsenhausen problem, respectively, when the transmission noise intensity goes to zero or infinity. Assuming a small transmission noise intensity, an asymptotic approach is then used in order to find an approximated cost. A necessary condition for asymptotically optimal strategies has been obtained using a variational approach and it is shown that the linear strategies, with slightly different coefficients than the noiseless transmission case, satisfy the necessary condition.

1.2 Application to Power Allocation in Cellular Radio Networks

A distributed Dynamic Channel and Power Allocation (DCPA) scheme based on a novel predictive power control algorithm is proposed [2, Appendix B]. Power control is considered an efficient scheme to mitigate co-channel and multiple-access interference in cellular radio systems. Various approaches have been proposed in recent years to design power control algorithms. We focus on the feedback algorithms that are based on Signal to Interference plus Noise Ratios (SIR-based algorithms). We review SIR threshold approaches and then discuss how power control design can be formulated as a decentralized regulation problem. We use a robust control framework to analyze global stability of a network on a single channel. We obtain a sufficient condition, which guarantees that the deviations of the power levels form their optimal values remain bounded, even when the channel gains change, as long as the network stays feasible [3, Appendix C]. The Minimum Interference Dynamic Channel Assignment algorithm is employed, while simple Kalman Filters are designed to provide the predicted measurements of both the channel gains and the interference levels, which are then used to update the power levels. Extensive computer simulations are carried out to show the improvement in performance, under the dynamics of user arrivals and departures and user mobility. It is shown that the number of dropped calls and the number of blocked calls are decreased while, on average, fewer channel reassignments per call are required [2, Appendix B].

1.3 Periodic Control

A Π test is presented for determining when a controller with periodic gains is superior to a LTI compensator for a class of LQ strong stabilization problems [4, Appendix D]. It has been noted that only strongly stabilizing compensators can stabilize a certain type of decentralized system. For systems with strictly proper transfer functions, it is proven that stable high frequency periodic controllers based on weak variations about the LTI

case cannot give better performance than stable LTI compensators. In the development, a means to evaluate the second partials of functions with respect to matrix valued parameters is introduced. These techniques can be trivially modified to deal with problems involving optimizing decentralized controllers for systems with fixed modes.

2. Fault Detection and Distributed Detection Filters

2.1 A Generalized Least-Squares Fault Detection Filter

A fault detection and identification algorithm is determined from a generalization of the least-squares derivation of the Kalman filter [5, Appendix E]. The objective of the filter is to monitor a single fault called the target fault and block other faults, which are called nuisance faults. The filter is derived from solving a min-max problem with a generalized least-squares cost criterion which explicitly makes the residual sensitive to the target fault, but insensitive to the nuisance faults. It is shown that this filter approximates the properties of the classical least-squares fault detection filter such that in the limit where the weighting on the nuisance fault is zero, the generalized least-squares fault detection filter becomes equivalent to the unknown input observer where there exists a reduced-order filter. Filter designs can be obtained for both linear time-invariant and time-varying systems.

2.2 Robust Multiple-Fault Detection Filter

A new robust multiple-fault detection and identification algorithm is proposed [6, Appendix F]. Different from other algorithms which explicitly force the geometric structure by using eigenstructure assignment or geometric theory, this algorithm is derived by solving an optimization problem. The output error is divided into several subspaces. For each subspace, the transmission from one fault, denoted the associated target fault, is maximized, and the transmission from other faults, denoted the associated nuisance fault, is minimized. Therefore, each projected residual of the robust multiple-fault detection filter is affected primarily by one fault and minimally by the other faults. The transmission from process and sensor noise is also minimized so that the filter is robust with respect to these disturbances. It is shown that this filter approximates the properties of the restricted diagonal filter of which the Beard-Jones detection filter is a special case. In the limit where the weighting on each associated nuisance fault transmission goes to infinity, the geometric structure of the restricted diagonal detection filter is recovered. When it is not in the limit, the filter only isolates the faults within approximate invariant subspaces. This new feature allows the filter to be potentially more robust since the filter structure is less constrained. Filter design can be obtained for both time- invariant and time-varying linear systems.

2.2 Optimal Stochastic Fault Detection Filter

A fault detection and identification algorithm, called optimal stochastic fault detection filter, is determined [7, Appendix G]. The objective of the filter is to monitor a single fault called the target fault and block other faults, which are called the nuisance faults in

the presence of the process and sensor noises. The filter is derived by maximizing the transmission from the target fault to the projected output error while minimizing the transmission from the nuisance faults. Therefore, the residual is affected primarily by the target fault and minimally by the nuisance faults. The transmission from the process and sensor noises is also minimized so that the filter is robust with respect to these disturbances. This filter is a special case of the detection filter of [6, Appendix F]. It is shown that this filter approximates the properties of the classical fault detection filter such that in the limit where the weighting on the nuisance fault transmission goes to infinity, the optimal stochastic fault detection filter becomes equivalent to the unknown input observer. However, the nuisance fault directions and their associated invariant zero directions must be included in the invariant subspace generated by the optimal stochastic fault detection filter. The asymptotic behavior of the filter as the weighting on the nuisance fault transmission becomes large is determined by using a perturbation method and it is shown that the geometric structure of the unknown input observer is recovered. Filter designs can be obtained for both time-invariant and time-varying systems.

2.3 Fault Reconstruction from Sensor and Actuator Failures

An approach for reconstructing sensor and actuator faults from the residual is proposed [8, Appendix H]. The transfer matrix from the faults to the residual is derived in terms of the eigenvalues of the fault detection filter associated with the invariant subspaces of the fault and the invariant zeros of the faults. For each fault, all possible fault reconstruction processes are derived parameterized by applying a projector to the transfer matrix and taking its inverse. Then, the optimal fault reconstruction process is determined by minimizing the ratio of the H_2 norm of the projected transfer matrix from the disturbance to the H_2 norm of the projected transfer matrix from the fault. For the existence of the fault reconstruction process, the invariant zeros of the fault have to be in the left half plane. Furthermore, for reconstructing a sensor fault, the system has to be detectable with respect to the other sensors.

2.4 A Decentralized Fault Detection Filter

The decentralized fault detection filter has a structure that results from merging decentralized estimation theory with the game theoretic fault detection filter [9, Appendix I]. A decentralized approach may be the ideal way to health monitor large-scale systems, since it decomposes the problem down into (potentially smaller) "local" problems. These local results are then blended into a "global" result that describes the health of the entire system. The benefits of such an approach include added fault tolerance and easy scalability. An example given at the end of the paper demonstrates the use of this filter for a platoon of cars proposed for an advanced vehicle control system.

2.5 Application of Detection Methods to Target Association

A residual-based scheme for solving the radar track association problem using bearings-only measurements is developed [10, Appendix J]. To accomplish track association between two stations, we analyze the residuals of a bank of nonlinear filters called modified gain extended Kalman filters (MGEKFs). Once tracks have been associated between two stations, tracks from additional stations may be associated with tracks from the first two stations by checking algebraic parity equations. Traditional track association

methods rely on the local stations' estimated target positions and error variances, which may be quite inaccurate when using bearings-only measurements. Our method bypasses this difficulty, since our filters use raw data from two stations. An example illustrates the effectiveness of our methods.

3 Nonlinear Estimation Applied to Relative GPS Navigation

3.1 Exact Statistical Solution of Pseudorange Equations

Although the exact GPS solution proposed by Bancroft is nonlinear, it may be manipulated into a linear form when 5 or more satellites are visible [11, Appendix K]. This linear form is exact, as opposed to the linear solution obtained via repeated linearization in the iterated least squares (ILS) method. By virtue of this exactness, the solution of the linear form is always the true user position, while the ILS may converge to an incorrect solution (this is especially common when the GPS user is in space).

When the measured pseudoranges are noisy, the linear structure ensures that the position estimate will converge to the correct value and that the error covariance of the estimate is known, guarantees that have not been found for nonlinear estimators that use the Bancroft solution directly. The conversion to the linear form excludes information present in a single scalar nonlinear measurement equation. We demonstrate several procedures for refining the linear estimate with this remaining information. In addition, we show that the methodology developed for direct GPS solutions can be applied to create linear direct methods for differential GPS problems.

3.2 Multiple Hypothesis Sequential Probability Ratio Tests for Resolving Integer Ambiguity in GPS

Two statistical techniques appropriate for the "validation" of integer ambiguities and the detection of cycle slips are developed [12, appendix L]. The multiple hypothesis Wald sequential probability ratio test (SPRT) can find the conditional probability that each set of integer biases under consideration is the true bias condition. The multiple hypothesis Shiryayev SPRT determines the conditional probability that the integer biases have jumped from the nominal bias condition to each member of a collection of other bias conditions. Hence, the Wald SPRT is a method for validating the integer ambiguities during the initial ambiguity resolution, while the Shiryayev SPRT can be used to monitor for cycle slips. Each of these multiple hypothesis SPRTs (MHSPRTs) makes use of two measurement residuals. One is geometric combination of the carrier phase measurements, and the other is generated by differencing the carrier phase measurements with code measurements.

Prior work on cycle slip monitoring has focused solely on the detection of the occurrence of a cycle slip in the fastest time, balanced against the probability of issuing a false alarm. Once a disruption has occurred, the ambiguity resolution process must restart from scratch. The Shiryayev SPRT bypasses this problem, as it announces the location of the biases after the jump, in addition to the time of the cycle slip. The calculations for the MHSPRTs are not linked to any particular distribution, unlike prior efforts. Only the probability density functions of the measurement residuals are required. Hence, the

techniques can correctly compensate for non-Gaussian errors in measurement such as multipath. For each hypothesis under consideration, the MHSPRTs yield the probability of that hypothesis being the correct one. The "state" of the MHSPRT recursions is the vector of all of these probabilities. Information from past measurements is embedded in this state. This recursive, probabilistic framework makes it very straightforward to add new hypotheses into the set of possible bias conditions while retaining information from prior measurements. In contrast, there is no way to do this for other techniques, since they are based on cumulative sums. Results from successful simulations and field experiments show the efficacy of these techniques.

Publications

[1] Kambiz Shoarinejad, Jason L. Speyer, and Ioannis Kanellakopoulos, "A Stochastic Decentralized Control Problem with Noisy Communication," *SIAM Journal of Control Optim., Vol. 41, No.3, 2002, pp. 975-990.*

[2] Kambiz Shoarinejad, Jason L. Speyer, and Gregory J. Pottie, "A distributed scheme for integrated predictive dynamic channel and power allocation in cellular radio networks" Proceedings of the IEEE Globecom Conference 2001 and to be publish in the *IEEE Transactions on Wireless Communication (IEEE ToWC).*

.

[3] Kambiz Shoarinejad, Jason L. Speyer, Fernando Paganini, and Gregory J. Pottie, "Global Stability of Feedback Power Control Algorithms for Cellular Radio Networks" Proceedings of the IEEE CDC'01.

[4] Jonathan D. Wolfe and Jason L. Speyer, "The periodic optimality of LQ controllers satisfying strong stabilization," Proceedings of the IFAC workshop on periodic control, August, 2001 and to be publish in *Automatica.*

[5] Robert H. Chen and Jason L. Speyer, "A generalized least-squares fault detection filter," *International Journal of Adaptive Control and Signal Processing*, vol. 14, pp. 747-757, 2000

[6] Robert H. Chen and Jason L. Speyer, "Robust Multiple-Fault Detection Filter," The special issue of condition monitoring, fault detection and isolation in the *International Journal of Robust and Nonlinear Control*, Vol. 12, Issue 8, 2002.

[7] Robert H. Chen, D. Lewis Mingori and Jason L. Speyer, "Optimal Stochastic Fault Detection Filter," *Automatica,* vol. 39 (2003) 377-390.

[8] Robert H. Chen and Jason L. Speyer, "Fault Reconstruction from Sensor and Actuator Failures," Proceedings of the IEEE Conference on Decision and Control, December, 2001

[9] Walter H. Chung, Jason L. Speyer and Robert H. Chen, "A Decentralized Fault Detection Filter," *ASME J. of Dynamic Systems, Measurement, Control*, Vol. 123, 2001.

[10] Jonathan D. Wolfe and Jason L. Speyer, "Target Association Using Detection Methods," *AIAA J. Guidance, Control, and Dynamics,* Vol. 25, No. 6, November-December, 2002.

[11] Jonathan D. Wolfe and Jason L. Speyer, "Exact Statistical Solution of Pseudorange Equations" Proceedings of the ION GPS 2001 and to be published in *The Journal of the Institute of Navigation*

[12] Jonathan D. Wolfe, Jason L. Speyer, and Walton R. Williamson, "Multiple Hypothesis Sequential Probability Ratio Tests for Resolving Integer Ambiguity in GPS," Proceedings of the ION GPS 2001 and to be published in *The Journal of the Institute of Navigation*

Personnel Supported
Graduate Students: Robert Chen, Jonathan Wolfe, Kambiz Shoarinejad, and Charles Dillon, Frederico Najson.

Interactions
Air Vehicle Division,WPAFB, Contact:Dr. Siva Banda, Munitions Division, Eglin AFB, Contact Johny Evers and Rob Murphy

Transitions
Nasa Dryden Flight Research Center, Integer ambiguity resolution, Contact: Gerard Schkolnik(661-258-3681)

Advisory Function
Member of the Air Force Scientific Advisory Board

Honors/Awards
Fellow of the IEEE and AIAA. Recipient of the IEEE Third Millennium Medal
Department of the Air Force Award for Meritorious Civilian Service, 2001
NASA Public Service Group Achievement Award awarded to the UCLA Autonomous Vehicles System Instrumentation Laboratory

Appendix A

"A Stochastic Decentralized Control Problem with Noisy Communication,"

Kambiz Shoarinejad, Jason L. Speyer, and Ioannis Kanellakopoulos,

SIAM Journal of Control Optim., Vol. 41, No.3, 2002, pp. 975-990.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 975–990

A STOCHASTIC DECENTRALIZED CONTROL PROBLEM WITH NOISY COMMUNICATION*

KAMBIZ SHOARINEJAD†, JASON L. SPEYER‡, AND IOANNIS KANELLAKOPOULOS§

Abstract. A simple decentralized stochastic control problem is considered where the nonclassical nature of the information pattern is induced by the uncertainty on the information transmission in the system. This is, in fact, a reformulation of the Witsenhausen counterexample, where the first station is allowed to send its information to the second station through a noisy channel. Nonconvexity of the problem in this new formulation has been established, and it is shown how this formulation relates to a classical problem and the Witsenhausen problem, respectively, when the transmission noise intensity goes to zero or infinity. Assuming small transmission noise intensity, we then use an asymptotic approach in order to find an approximated cost. A necessary condition for asymptotically optimal strategies has been obtained using a variational approach, and it is shown that the linear strategies, with slightly different coefficients than the noiseless transmission case, satisfy the necessary condition.

Key words. optimal stochastic control, decentralized systems, asymptotic analysis

AMS subject classifications. 93E20, 93A14

PII. S0363012901385629

1. Introduction.

Coordinating and controlling dynamic systems in spatial networks has always been a challenging problem for system designers. It is now attracting more attention as various new applications are emerging in a very wide range from autonomous vehicles in formation to flow and congestion control in computer networks. However, there are still some major difficulties in dealing with such systems. The main characteristics of any decentralized system is that the information is distributed among different stations and the performance of the system depends highly on the corresponding information pattern, i.e., *who knows what and when*. The stations may communicate with each other possibly by signaling through noisy channels. Even though there might be some physical constraints on the information structure of the system (e.g., locations of the sensors, the actuators, and the transmitters), in general, an optimal information pattern should be obtained. Then, based on the locally available information, a set of coordinated local strategies should be designed in order to achieve a common objective. In many cases, however, we will end up with nonconvex functional optimization problems, which are usually very difficult to solve.

One such class of problems is when a decentralized system has a *nonclassical* information pattern which is not partially nested. The information pattern is called nonclassical when the distributed stations do not have access to the same information and/or some stations do not have perfect recall (i.e., they lose information). Moreover,

*Received by the editors February 26, 2001; accepted for publication (in revised form) March 2, 2002; published electronically September 19, 2002. The results in this paper were partially presented at the 38th IEEE Conference on Decision and Control and the 1999 American Control Conference. This research was supported in part by the National Science Foundation under grant ECS-9502945, the Air Force Office of Scientific Research under grant F49620-97-1-0272, and the Office of Naval Research under award N00014-97-1-0939.

http://www.siam.org/journals/sicon/41-3/38562.html

†Innovics Wireless Inc., 11500 Olympic Boulevard, Suite 398, Los Angeles, CA 90064 (kambiz@innovics.wireless.com).

‡UCLA Mechanical and Aerospace Engineering, Box 951597, Los Angeles, CA 90095-1597 (speyer@seas.ucla.edu).

§Voyan Technology, 3255-7 Scott Boulevard, Santa Clara, CA 95054 (ioannis@voyan.com).

a nonclassical information pattern is not partially nested when some stations cannot reconstruct the previous actions of other stations which have affected their own local information. Unfortunately, this happens in many decentralized systems.

In 1968, Witsenhausen provided a simple example in [1] in which there are only two stations, the dynamics are linear, the underlying uncertainties are additive and Gaussian, and the cost is quadratic. The information pattern, however, is nonclassical. This example motivated much research on the links between decentralized stochastic control problems and team theory and the effects of different information patterns on decentralized systems. Although it is a very simple example, it demonstrates the main difficulties induced by nonclassical information patterns. In this example, one station acts first and affects the information available to the next station, while there is no way for the second station to determine the action of the first station. The existence of the optimal design was established in [1], where a nonlinear set of strategies was also proposed which showed that no affine strategy could be optimal.

This seemingly simple example, which is also called Witsenhausen's counterexample, turned out to be extremely hard. It is still outstanding after more than 30 years. It was later shown in [2] that when the uncertainty on the information available to the first station is small, linear strategies would still be optimal over a large class of nonlinear strategies. Intuitively, when the uncertainty on the information of the first station is small, the second station will also be able to *guess* what that information was. Therefore, since the problem is cooperative in the sense that the stations are aware of each others' strategies, the second station can almost reconstruct the action of the first station, and there is no need for any kind of signaling among the stations through the dynamics of the system. In Witsenhausen's problem, the nonclassical nature of the information pattern is a result of the fact that the information available to the first station is completely inaccessible for the second station. However, recent advances in computing and communication technologies make it possible for the stations in many decentralized systems to communicate different pieces of information. But communications can never be perfect, and there is always some uncertainty involved. Unfortunately, such uncertainty will again induce a nonclassical nature on the information pattern of the system.

In this paper, we reformulate Witsenhausen's problem by allowing the first station to communicate its information with the second station through a noisy channel. Then we show that as long as there is noise in the transmission, the main difficulties will persist. Specifically, the cost might still be nonconvex with respect to the strategies. We then consider the two limit cases where the transmission uncertainty becomes either very large or negligible. We show how this new formulation covers a wide range of problems, from the classical linear quadratic Gaussian (LQG) problem to the Witsenhausen counterexample.

When the transmission noise intensity is small, one would expect the optimal strategies to be very close to the corresponding strategies for the noiseless transmission case. Our next objective in this paper is to investigate this case through an asymptotic analysis.

In section 2, we present the problem formulation. In section 3, we obtain an alternative form for the performance index, which clearly shows the possible nonconvexity of the cost with respect to the strategies. In section 4, we consider the two limit cases, i.e., when the transmission noise intensity goes to zero or infinity. In section 5 we assume a small uncertainty on the transmission and approximate the cost by expanding it in terms of the small transmission noise intensity. In section 6, we use a variational approach in order to find a necessary condition for the strategies that

minimize the approximated cost. As we shall see, we will actually have a singular optimization problem. We then show that the asymptotically optimal strategies can still be linear with slightly different coefficients than the corresponding strategies for the noiseless transmission case. We provide concluding remarks in the final section.

2. Problem description. Consider a two-stage stochastic problem with the following state equations:

$$x_1 = x_0 + u_1, \tag{2.1}$$

$$x_2 = x_1 - u_2, \tag{2.2}$$

where x_0 is the initial state, which is assumed to be a zero mean Gaussian random variable with variance σ_0^2. The information pattern of the system is specified by the following output equations:

$$z_1 = x_0, \tag{2.3}$$

$$z_2 = \begin{bmatrix} x_0 + v_t \\ x_0 + u_1 + v_2 \end{bmatrix} := \begin{bmatrix} z_{21} \\ z_{22} \end{bmatrix}, \tag{2.4}$$

where v_2 is the measurement noise for the second station, which is also assumed to be a zero mean Gaussian random variable with unit variance. As we can see, the information available to the first station is being transmitted to the second station, and the communication uncertainty is modeled by an additive Gaussian noise $v_t \sim \mathcal{N}(0, \epsilon^2)$. Also, x_0, v_2, and v_t are all assumed to be independent of each other. It is clear that we have simply modeled the received information signal as the transmitted signal plus the Gaussian transmission noise. While this model can be quite realistic for analog communication systems, it may not be well justified when digital communication is used. In digital communication systems the signal is quantized, coded, and sent through the channel. Still, the channel noise may realistically be assumed to be additive and Gaussian, but sophisticated modulation and coding schemes make it difficult to assume a simple additive Gaussian uncertainty for the received *information signal.* However, if we try to incorporate the quantization effects along with the bit error probability distribution for some *good* coding and modulation schemes in order to model the communication uncertainties, we will end up with models which could still be approximated, to some degree, by simple additive Gaussian models. Moreover, since there are already major difficulties in dealing with decentralized nonclassical information patterns, using more complex models for communication uncertainties may not seem very reasonable at this point. Furthermore, we believe that the results obtained under such a simplifying assumption would still serve as a guideline for finding the true nature of the optimal decentralized strategies. The objective is now to design the control strategies γ_1 and γ_2,

$$u_1 = \gamma_1(z_1), \tag{2.5}$$

$$u_2 = \gamma_2(z_2), \tag{2.6}$$

in order to minimize the cost function

$$J = E\left[k^2 u_1^2 + x_2^2\right], \tag{2.7}$$

where $k^2 > 0$ is a given constant. Note that this is a sequential stochastic control problem in the sense that the second station acts after the first station. In other words,

$$-\dot{\bar{S}} = \bar{S}[A_{11} - \Gamma_{11} - G_1(D_2^{\mathrm{T}}C_2^{\mathrm{T}}V^{-1}C_2D_2)^{-1}D_2^{\mathrm{T}}C_2^{\mathrm{T}}V^{-1}C_1]$$
$$+ [A_{11} - \Gamma_{11} - G_1(D_2^{\mathrm{T}}C_2^{\mathrm{T}}V^{-1}C_2D_2)^{-1}D_2^{\mathrm{T}}C_2^{\mathrm{T}}V^{-1}C_1]^{\mathrm{T}}\bar{S}$$
$$+ \bar{S}[-N_1Q_1N_1^{\mathrm{T}} + G_1(D_2^{\mathrm{T}}C_2^{\mathrm{T}}V^{-1}C_2D_2)^{-1}G_1^{\mathrm{T}}]\bar{S} - C_1^{\mathrm{T}}\bar{H}^{\mathrm{T}}V^{-1}\bar{H}C_1 \quad (23)$$

By substituting (21) and (22) into (20a), the reduced-order limiting generalized least-squares fault detection filter is

$$\dot{\hat{\eta}}_1 = (A_{11} - \Gamma_{11})\hat{\eta}_1 + M_1u + [G_1(D_2^{\mathrm{T}}C_2^{\mathrm{T}}V^{-1}C_2D_2)^{-1}D_2^{\mathrm{T}}C_2^{\mathrm{T}}V^{-1} + \bar{S}^{-1}C_1^{\mathrm{T}}\bar{H}^{\mathrm{T}}V^{-1}\bar{H}](y - C_1\hat{\eta}_1) \quad (24)$$

Note that Γ_{11} can be computed *a priori*. In the limit, the residual (3) becomes

$$r = \hat{H}(y - C_1\hat{\eta}_1) \quad (25)$$

because $\hat{H}C_2 = 0$ from (21) and $\mathrm{Ker}\,\hat{H} = \mathrm{Ker}\,\bar{H}$.

7. EXAMPLE

In this section, two numerical examples are used to demonstrate the performance of the generalized least-squares fault detection filter. In Section 7.1, the filter is applied to a time-invariant system. In Section 7.2, the filter is applied to a time-varying system.

7.1. Example 1

In this section, two cases for a time-invariant problem are presented. The first one shows that the sensitivity of the filter (8) to the nuisance fault decreases when γ is smaller. The second one shows that the sensitivity of the reduced-order limiting filter (24) to the target fault increases when Q_1 is larger. The system matrices are

$$A = \begin{bmatrix} 0 & 3 & 4 \\ 1 & 2 & 3 \\ 0 & 2 & 5 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad F_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad F_2 = \begin{bmatrix} 5 \\ 1 \\ 1 \end{bmatrix}$$

In the first case, the steady-state solutions to the Riccati equation (9) are obtained with weightings chosen as $Q_1 = 1$, $Q_2 = 1$, and $V = I$ when $\gamma = 10^{-4}$ and 10^{-6}, respectively. The top two figures of Figure 1 show the frequency response from both faults to the residual (3). The left one is $\gamma = 10^{-4}$, and the right one is $\gamma = 10^{-6}$. The solid lines represent the target fault, and the dashed lines represent the nuisance fault. This example shows that the nuisance fault transmission can be reduced by using a smaller γ while the target fault transmission is not affected.

In the second case, the steady-state solutions to the reduced-order limiting Riccati equation (23) are obtained with $V = 10^{-4}I$ when $Q_1 = 0$ and 0.0019, respectively. The lower two figures of Figure 1 show the frequency response from the target fault and sensor noise to the residual (25). The left one is $Q_1 = 0$, and the right one is $Q_1 = 0.0019$. The solid lines represent the target fault, and the dashed lines represent the sensor noise. This example shows that the

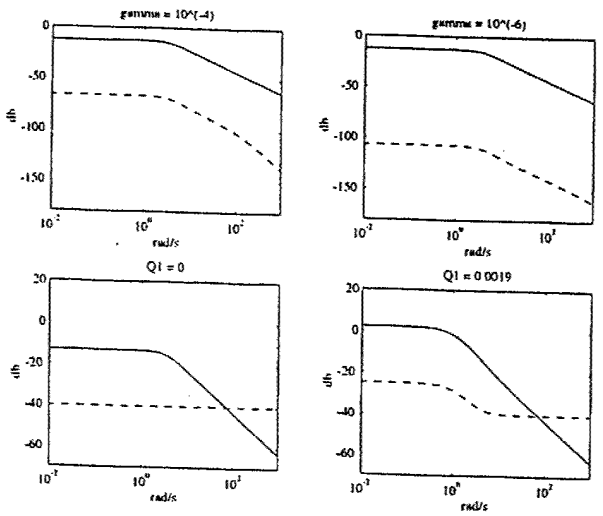


Figure 1. Frequency response of the residual.

sensitivity of the filter to the target fault can be enhanced by using a larger Q_1. The sensor noise transmission also increases because part of the sensor noise comes through the same direction as the target fault. However, the sensor noise transmission is small compared to the target fault transmission. In this case, the nuisance fault transmission stays zero and is not shown in these figures. Note that when $Q_1 = 0$, the generalized least-squares fault detection filter is similar to Reference [2] which does not enhance the target fault transmission.

7.2. *Example 2*

In this section, the filter (8) and the reduced-order limiting filter (24) are applied to a time-varying system which is from modifying the time-invariant system in the previous section by adding some time-varying elements to A and F_2 matrices while C and F_1 matrices are the same:

$$A = \begin{bmatrix} -\cos t & 3 + 2\sin t & 4 \\ 1 & 2 & 3 - 2\cos t \\ 5\sin t & 2 & 5 + 3\cos t \end{bmatrix}, \quad F_2 = \begin{bmatrix} 5 - 2\cos t \\ 1 \\ 1 + \sin t \end{bmatrix}$$

The Riccati equation (9) is solved with $Q_1 = 1$, $Q_2 = 1$, $V = I$ and $\gamma = 10^{-5}$ for $t \in [0,\ 25]$. The reduced-order limiting Riccati equation (23) is solved with the same Q_1 and V. Figure 2 shows the time response of the norm of the residuals when there is no fault, a target fault and a nuisance fault, respectively. The faults are unit steps that occur at the fifth second. In each case, there is no sensor noise. The left three figures show the residual (3) for the filter (8). There is a small nuisance fault transmission because (8) is an approximate unknown input observer. The right three figures show the residual (25) for the reduced-order limiting filter (24). Note that the nuisance fault transmission is zero. This example shows that both filters, (8) and (24), work well for time-varying systems.

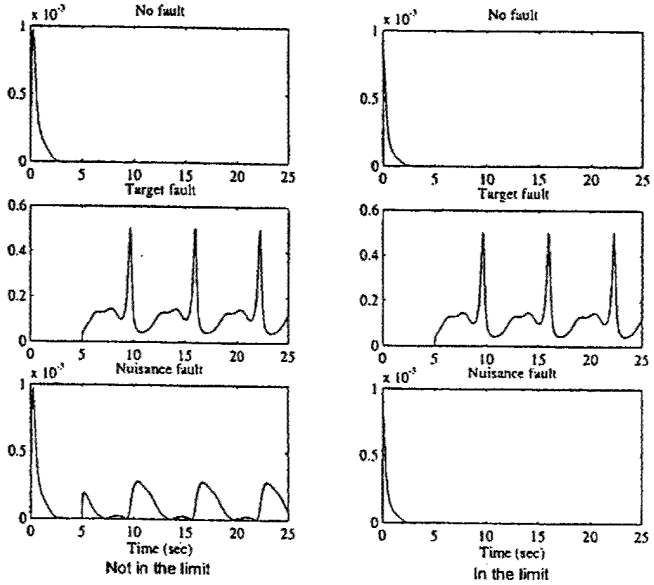


Figure 2. Time response of the residual.

8. CONCLUSION

The generalized least-squares fault detection filter is derived from solving a min–max problem which makes the residual sensitive to the target fault, but insensitive to the nuisance faults. In the limit where the weighting on the nuisance faults is zero, the filter becomes equivalent to the unknown input observer which places the nuisance faults into a minimal (C, A)-unobservability subspace and there exists a reduced-order filter. Since the target fault is explicit in the problem formulation, the sensitivity of the filter to the target fault can be enhanced. Filter designs can be obtained for both linear-time-invariant and time-varying systems.

ACKNOWLEDGEMENTS

This work was sponsored by Air Force Office of Scientific Research, Award No. F49620-97-1-0272 and by the California Department of Transportation, Agreement No. 65A0013, MOU315.

REFERENCES

1. Massoumnia M-A, Verghese GC, Willsky AS. Failure detection and identification. *IEEE Transactions on Automatic Control* 1989; **AC-34**(3):316–321.
2. Chung WH, Speyer JL. A game theoretic fault detection filter. *IEEE Transactions on Automatic Control* 1998; **AC-43**(2):143–161.
3. Chen RH, Speyer JL. Optimal stochastic fault detection filter. *Proceedings of the American Control Conference*, 1999; 91–96.
4. Bryson AE, Ho Y-C. *Applied Optimal Control: Optimization, Estimation and Control*. Hemisphere: Washington, DC, 1975.
5. Bell DJ, Jacobsen DH. *Singular Optimal Control Problems. Mathematics in Science and Engineering*, vol. 117. Academic Press: New York, 1975.
6. Chen RH. Fault detection filters for robust analytical redundancy. *PhD Thesis*, University of California, Los Angeles, 2000.

the order in which the stations apply their control actions does not depend on the uncertainties in the system. We see that the first controller has perfect information but its action is costly. In contrast, the second controller has inexpensive control but noisy information. Since the second station does not know what the first station knew, due to the transmission noise, we do not have perfect recall, and hence we still have a nonclassical pattern. If there was no transmission noise, we would have a classical information pattern for which the unique optimal strategies are known to be linear in the information.

3. An alternative form for the performance index. In this section, we show how the performance index may be expressed in terms of the Fisher information matrix, which indicates that the cost may not be convex in the strategies.

For simplicity, and similarly to the Witsenhausen problem, we define

$$f(z_1) := z_1 + \gamma_1(z_1) = x_0 + u_1, \tag{3.1}$$

$$g(z_2) := \gamma_2(z_2) = u_2. \tag{3.2}$$

Then the cost can be expressed as

$$\begin{aligned} J &= E\left[k^2 u_1^2 + x_2^2\right] \\ &= E\left[k^2 (z_1 - f(z_1))^2 + (f(z_1) - g(z_2))^2\right] \\ &:= J(f,g). \end{aligned} \tag{3.3}$$

If we fix the function f, the optimal strategy g will clearly be obtained as the conditional expectation, i.e.,

$$g^*(z_2) = \arg\min_g J(f,g) = E[f(z_1)\,|z_2]. \tag{3.4}$$

Substituting the above equation back in the cost, we get

$$\begin{aligned} J^*(f) &:= J(f,g^*) \\ &= k^2 E\left[(z_1 - f(z_1))^2\right] + E\left[(f(z_1) - g^*(z_2))^2\right] \\ &= k^2 E\left[(z_1 - f(z_1))^2\right] + E\left[(f(z_1))^2\right] - E\left[(g^*(z_2))^2\right], \end{aligned} \tag{3.5}$$

where we have used the orthogonality property of the conditional expectation

$$E[(f(z_1) - g^*(z_2))\, g^*(z_2)] = 0. \tag{3.6}$$

It is important to note the minus sign in the third term in (3.5). As we shall see, this minus sign could indeed destroy the convexity of the cost with respect to the strategies.

The objective is now to express the cost $J^*(f)$ in terms of only one strategy f. In doing so, we use the following lemma, which shows how $g^*(z_2)$ may be expressed in terms of information z_2 and its probability density function.

LEMMA 3.1. *The optimal strategy $g^*(z_2)$ can be expressed as*

$$g^*(z_2) = z_{22} + \frac{\partial}{\partial z_{22}} \ln p(z_2), \tag{3.7}$$

where $p(z_2) = p(z_{21}, z_{22})$ is the probability density function for the information available to the second station.

Proof. We have

$$g^*(z_2) = \int f(z_1)\, p(z_1 | z_2)\, dz_1$$

$$= \frac{\int f(z_1)\, p(z_1, z_2)\, dz_1}{\int p(z_1, z_2)\, dz_1}, \tag{3.8}$$

where $p(z_1, z_2)$ is the joint probability density of z_1 and z_2. At the same time, one can write

$$f(z_1)\, p(z_1, z_2) = z_{22} p(z_1, z_2) + \frac{\partial}{\partial z_{22}} p(z_1, z_2). \tag{3.9}$$

This can be shown as

$$\begin{aligned} z_{22} p(z_1, z_2) + \frac{\partial}{\partial z_{22}} p(z_1, z_2) &= z_{22} p(z_1, z_2) + \frac{\partial}{\partial z_{22}} p(z_2 | z_1)\, p(z_1) \\ &= z_{22} p(z_1, z_2) + \frac{\partial}{\partial z_{22}} p_{(v_t, v_2)} \left(\begin{bmatrix} z_{21} \\ z_{22} \end{bmatrix} - \begin{bmatrix} z_1 \\ f(z_1) \end{bmatrix} \right) p(z_1) \\ &= z_{22} p(z_1, z_2) + \frac{\partial}{\partial z_{22}} \left(\frac{1}{2\pi\epsilon} \exp\left(-\frac{(z_{21} - z_1)^2}{2\epsilon^2} - \frac{(z_{22} - f(z_1))^2}{2} \right) \right) p(z_1) \\ &= f(z_1)\, p(z_1, z_2), \end{aligned} \tag{3.10}$$

where we have used the specific form of the information available to the second station and the fact that $v_t \sim \mathcal{N}(0, \epsilon^2)$ and $v_2 \sim \mathcal{N}(0, 1)$ are independent. By substituting for $f(z_1)\, p(z_1, z_2)$ from (3.9) back in (3.8) and integrating with respect to z_1, the expression in (3.7) is obtained. □

As we shall see, when we try to express the performance index in terms of only a single strategy f, a *Fisher information* term comes up in the cost. Fisher information is originally obtained in the Cramer–Rao bound, which is a measure for the minimum error in estimating a parameter based on the value of a random variable. However, by introducing a location parameter, an alternative form of the Fisher information may be defined for a random variable with a given distribution. This alternative form is, in fact, related to the entropy measure (see [3, p. 494]). We first present the definition for the Fisher information matrix.

DEFINITION 3.2. *The Fisher information matrix for a random vector Z is defined as*

$$I_f(Z) := E\left[\nabla_z^T \ln p(z) \cdot \nabla_z \ln p(z) \right], \tag{3.11}$$

where $p(z)$ is the probability density function for the random variable Z and ∇_z denotes the gradient vector with respect to z:

$$\nabla_z := \left[\frac{\partial}{\partial z_1} \cdots \frac{\partial}{\partial z_n} \right], \tag{3.12}$$

where z_i is the ith component in the random vector.

We are now ready to present the alternative expression for the performance index.

THEOREM 3.3. *The performance index* (3.5) *can be written as*

$$J^*(f) = k^2 E\left[(z_1 - f(z_1))^2 \right] + 1 - I_f(Z_2)_{22}, \tag{3.13}$$

where $I_f\,(Z_2)_{22}$ *is, in fact, the* $(2,2)$ *element of the* Fisher information matrix *for the random vector* Z_2. *The subscript* f *indicates the fact that it actually depends on the form of the strategy* f, *which is present in the definition of* z_2 *and would affect its probability density function.*

Proof. Using (3.7), we first obtain $E[(g^*(z_2))^2]$. We have

$$E\left[z_{22}^2\right] = E\left[\left(f\left(z_1\right)\right)^2\right] + 1, \tag{3.14}$$

and

$$E\left[z_{22}\frac{\partial}{\partial z_{22}}\ln p\left(z_2\right)\right] = \int\!\!\int_{-\infty}^{+\infty} z_{22}\frac{\partial}{\partial z_{22}}\ln\left(p\left(z_{21},z_{22}\right)\right)p\left(z_{21},z_{22}\right)dz_{21}dz_{22}. \tag{3.15}$$

If we integrate by parts with respect to z_{22}, we get

$$\begin{aligned}\int_{-\infty}^{+\infty} z_{22}\frac{\partial}{\partial z_{22}}\ln\left(p\left(z_{21},z_{22}\right)\right)p\left(z_{21},z_{22}\right)dz_{22} &= z_{22}p\left(z_{21},z_{22}\right)\big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} p\left(z_{21},z_{22}\right)dz_{22}\\ &= -p\left(z_{21}\right),\end{aligned} \tag{3.16}$$

where z_{22} is assumed to have a finite mean value, and therefore the first term becomes zero. Hence,

$$E\left[z_{22}\frac{\partial}{\partial z_{22}}\ln p\left(z_2\right)\right] = -1. \tag{3.17}$$

Therefore,

$$E\left[\left(g^*\left(z_2\right)\right)^2\right] = -1 + E\left[\left(f\left(z_1\right)\right)^2\right] + I_f\left(Z_2\right)_{22}, \tag{3.18}$$

where

$$I_f\left(Z_2\right)_{22} = E\left[\left(\frac{\partial}{\partial z_{22}}\ln p\left(z_2\right)\right)^2\right]. \tag{3.19}$$

Substituting (3.18) back in (3.5), we get (3.13) as an alternative form for representing the performance index. □

As we see, the cost is now expressed only in terms of one strategy f. Also, this somehow shows us that in order to minimize the cost, we need to get the lowest possible cost associated with the first station, while we transfer as much information as possible to the second station through the dynamics of the system. The possible nonconvexity of the cost with respect to f can also be seen from this alternative expression. It can be shown that the Fisher information term is a convex functional [4]. Therefore, $1 - I_f\left(Z_2\right)_{22}$ is concave and the sum of a convex and a concave functional may not be convex.

4. Limit cases. In this section we consider the two limit cases. First we consider the case where the transmission is noiseless, and then we investigate the case where the transmission noise intensity goes to infinity.

4.1. Noiseless transmission. Assume there is no uncertainty in transmitting information from the first to the second station, i.e., $\epsilon = 0$ and hence $z_{21} = z_1$. In this case, we have perfect recall and the information pattern is classical. We can write

$$\begin{aligned} p(z_2) &= p(z_{21}, z_{22}) = p(z_{22}|z_{21})\, p(z_{21}) \\ &= p(z_{22}|z_1)\, p(z_1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_{22} - f(z_1))^2}{2}\right) p(z_1). \end{aligned} \tag{4.1}$$

Then, from (3.7), we have

$$g^*(z_2) = f(z_1) = f(z_{21}), \tag{4.2}$$

which could directly be obtained from the original definition for g^*, i.e.,

$$g^*(z_2) = E[f(z_1)|\, z_2] = f(z_1), \tag{4.3}$$

because z_1 is exactly known when z_2 is given. Substituting this back in (3.5) and minimizing with respect to the strategy f, we have

$$g^*(z_2) = f(z_1) = z_1, \tag{4.4}$$

and hence

$$\gamma_1(z_1) = 0, \tag{4.5}$$

$$\gamma_2(z_2) = z_1, \tag{4.6}$$

which is the unique linear set of optimal strategies. This indeed turns out to be a very simple example of the well-known classical LQG problem.

4.2. Infinite transmission noise intensity. Another limit case is when the transmission noise intensity increases to infinity. In this case, z_{21} and z_{22} become independent and we have

$$p(z_2) = p(z_{21}, z_{22}) = p(z_{21})\, p(z_{22}). \tag{4.7}$$

The Fisher information term can now be written as

$$\begin{aligned} I_f(Z_2)_{22} &= \int\int_{-\infty}^{+\infty} \left(\frac{\partial}{\partial z_{22}} \ln p(z_{21}, z_{22})\right)^2 p(z_{21}, z_{22})\, dz_{21} dz_{22} \\ &= \int_{-\infty}^{+\infty} \left(\frac{\partial}{\partial z_{22}} \ln p(z_{22})\right)^2 p(z_{22})\, dz_{22} \\ &= I_f(Z_{22}), \end{aligned} \tag{4.8}$$

which is indeed the Fisher information content of z_{22} only. Hence,

$$J^*(f) = k^2 E\left[(z_1 - f(z_1))^2\right] + 1 - I_f(Z_{22}). \tag{4.9}$$

This is the same result that was presented for the Witsenhausen counterexample in [1]. Intuitively, when we have infinite transmission noise intensity, we might as well deny the access to z_1 for the second station, and this is exactly the case in Witsenhausen's counterexample. The optimal strategies for this case are still unknown. Witsenhausen

showed that the optimal solution exists, even if x_0 has a general distribution with a finite second moment [1]. He then showed that if one of the strategies is restricted to being affine, the other optimal strategy would also be affine. But then he provided a set of nonlinear strategies that could achieve a lower cost for some values of k^2 and σ_0.

Different approaches have been taken in order to find the optimal strategies. As mentioned before, an asymptotic approach was used in [2] for the case where σ_0 is small. More recently, in [5], [6], [7] it was shown how a neural network, trained by stochastic approximation techniques, can be employed as a nonlinear function approximator in order to approximate $f(z_1)$. It was demonstrated that the optimal $f^*(z_1)$ may not be strictly piecewise, as was suggested by Witsenhausen, but slightly sloped. Some researchers have tried to attack the problem numerically and use some sample and search techniques to find the solution. A discretized version of the problem was formulated in [8], which was later shown in [9] to be NP-complete and computationally intractable. It is recently asserted in [10] and [11] that a global optimum would be achieved by searching directly in the strategy space using the generalized step functions to approximate $f(z_1)$.

So far we have shown, through a simple example, how any uncertainty in the transmission of information between the stations in a distributed system can make the optimal control design very complicated and even intractable. Then, by considering the two limit cases, we showed how our example covers a very wide range of scenarios. Namely, we saw that for the noiseless transmission case, the unique optimal strategies, which are linear in the information, are easily obtained, whereas for the infinite transmission noise intensity, the optimal strategies are still unknown. Now a very feasible case to investigate is when the uncertainty on the information transmission is small. In fact, when the transmission noise intensity ϵ is small, one would still expect behavior similar to the noiseless transmission case for the optimal strategies. In the following sections, we consider this case. Namely, we assume a small intensity for v_t. Under this assumption, we obtain the first few terms in the expansion of the performance index in terms of ϵ. We then use the Hamiltonian approach in order to find a necessary condition for the strategies that minimize the approximated cost.

We show that the linear strategies, with slightly different coefficients than the corresponding coefficients for the noiseless transmission case, do indeed satisfy the necessary condition. This asymptotic analysis not only gives us insight on how the optimal strategies change as the transmission uncertainty is introduced but also provides us with a better sense of the complexities in the design procedure.

5. An expansion for the cost. Assume that the first station communicates with the second station through a low noise channel. In other words, the transmission noise intensity ϵ is assumed to be small. In this section, we will find an expansion for the cost in terms of ϵ. For this purpose, we first find an expansion for the probability density function of the information available to the second station, i.e., $p(z_2)$. Then we use (3.7) in order to find the corresponding expansion for $g^*(z_2)$. By substituting back in (3.5), we will obtain the expanded cost only in terms of f.

The probability density function for z_2 can be written as

$$p_\epsilon(z_2) := p(z_2) = \int_{-\infty}^{+\infty} p(z_{22}, z_{21}, z_1)\, dz_1 \tag{5.1}$$

$$= \int_{-\infty}^{+\infty} p(z_{22}|z_{21}, z_1)\, p(z_{21}|z_1)\, p(z_1)\, dz_1 \tag{5.2}$$

$$= \int_{-\infty}^{+\infty} p(z_{22}|z_1)\, p(z_{21}|z_1)\, p(z_1)\, dz_1 \tag{5.3}$$

$$= \int_{-\infty}^{+\infty} p(z_{22}|z_1)\, p_{v_t}(z_{21}-z_1)\, p(z_1)\, dz_1 \tag{5.4}$$

$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_{22}-f(z_1))^2}{2}\right) \frac{1}{\sqrt{2\pi\epsilon}} \exp\left(-\frac{(z_{21}-z_1)^2}{2\epsilon^2}\right) \times \frac{1}{\sqrt{2\pi\sigma_0}} \exp\left(-\frac{z_1^2}{2\sigma_0^2}\right) dz_1, \tag{5.5}$$

where for (5.3) we have used the facts that the σ-fields generated by $\{z_{21}, z_1\}$ and $\{z_1, v_t\}$ are the same and z_1, v_t, and v_2 are mutually independent. At this point, one should note that even though the joint probability density function $p(z_{22}, z_{21}, z_1)$ can be explicitly expressed as in (5.5), introduction into the performance index shows that determination of $f(z_1)$ still requires averaging over all random variables. This is another way of looking at the effect of a nonclassical information pattern, which is not partially nested. We therefore decide to follow an asymptotic approach.

For small ϵ, we now approximate $\ln p_\epsilon(z_2)$ by considering only the first three terms of its expansion around $\epsilon = 0$. Namely,

$$\ln p_\epsilon(z_2) \simeq \ln p_0(z_2) + \frac{\partial}{\partial\epsilon} \ln p_\epsilon(z_2)\Big|_{\epsilon=0} \epsilon + \frac{\partial^2}{\partial\epsilon^2} \ln p_\epsilon(z_2)\Big|_{\epsilon=0} \epsilon^2. \tag{5.6}$$

By making the change of variables

$$\epsilon y := z_1 - z_{21} \Rightarrow \epsilon dy = dz_1, \tag{5.7}$$

we can write $p_\epsilon(z_2)$ in the following form:

$$p_\epsilon(z_2) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_{22}-\bar{f}_\epsilon(y))^2}{2}\right) \frac{1}{\sqrt{2\pi\sigma_0}} \exp\left(-\frac{(z_{21}+\epsilon y)^2}{2\sigma_0^2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy, \tag{5.8}$$

where

$$\bar{f}_\epsilon(y) := f(\epsilon y + z_{21}). \tag{5.9}$$

It is now clear that

$$p_0(z_2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_{22}-f(z_{21}))^2}{2}\right) \frac{1}{\sqrt{2\pi\sigma_0}} \exp\left(-\frac{z_{21}^2}{2\sigma_0^2}\right), \tag{5.10}$$

and hence

$$\ln p_0(z_2) = -\frac{(z_{22}-f(z_{21}))^2}{2} - \frac{z_{21}^2}{2\sigma_0^2} + \ln\left(\frac{1}{2\pi\sigma_0}\right). \tag{5.11}$$

For the first order term, we have

$$\frac{\partial}{\partial\epsilon} \ln p_\epsilon(z_2)\Big|_{\epsilon=0} = \frac{1}{p_0(z_2)} \frac{\partial}{\partial\epsilon} p_\epsilon(z_2)\Big|_{\epsilon=0}. \tag{5.12}$$

On the other hand,

$$\begin{aligned}\frac{\partial}{\partial\epsilon}p_\epsilon(z_2)\bigg|_{\epsilon=0} &= \int_{-\infty}^{+\infty}\frac{\partial}{\partial\epsilon}\left\{\frac{1}{\sqrt{2\pi}}e^{-\frac{(z_{22}-\bar{f}_\epsilon(y))^2}{2}}\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{(z_{21}+\epsilon y)^2}{2\sigma_0^2}}\right\}\bigg|_{\epsilon=0}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy\\ &= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}(z_{22}-f(z_{21}))yf'(z_{21})e^{-\frac{(z_{22}-f(z_{21}))^2}{2}}\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{z_{21}^2}{2\sigma_0^2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy\\ &+\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{(z_{22}-f(z_{21}))^2}{2}}\frac{1}{\sqrt{2\pi}\sigma_0}\left(-\frac{z_{21}}{\sigma_0^2}\right)ye^{-\frac{z_{21}^2}{2\sigma_0^2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy=0.\end{aligned}$$

(5.13)

Therefore,

$$\frac{\partial}{\partial\epsilon}\ln p_\epsilon(z_2)\bigg|_{\epsilon=0}=0. \tag{5.14}$$

We could somehow expect this result. This is because we would expect the behavior of $p_\epsilon(z_2)$ to depend only on the variance of the Gaussian transmission noise, i.e., ϵ^2. Using (5.14), we can now obtain the second order term as

$$\frac{\partial^2}{\partial\epsilon^2}\ln p_\epsilon(z_2)\bigg|_{\epsilon=0}=\frac{1}{p_0(z_2)}\frac{\partial^2}{\partial\epsilon^2}p_\epsilon(z_2)\bigg|_{\epsilon=0}. \tag{5.15}$$

After some tedious but straightforward manipulations, we get

$$\begin{aligned}\frac{\partial^2}{\partial\epsilon^2}\ln p_\epsilon(z_2)\bigg|_{\epsilon=0} &= -f'^2(z_{21})+f''(z_{21})(z_{22}-f(z_{21}))+f'^2(z_{21})(z_{22}-f(z_{21}))^2\\ &+2f'(z_{21})(z_{22}-f(z_{21}))\left(-\frac{z_{21}}{\sigma_0^2}\right)-\frac{1}{2\sigma_0^2}+\frac{z_{21}^2}{\sigma_0^4}.\end{aligned} \tag{5.16}$$

We can now obtain a second order approximation for $\ln p_\epsilon(z_2)$ by substituting the corresponding terms from (5.11), (5.14), and (5.16) back into the expansion (5.6). In the next step, we substitute the expansion for $\ln p_\epsilon(z_2)$ in (3.7) in order to find the corresponding expansion for $g^*(z_2)$. Remember that $g^*(z_2)$ is the optimal strategy for the second station, assuming that the first station has a fixed strategy $\gamma_1(z_1)=f(z_1)-z_1$. We have

$$\begin{aligned}g^*(z_2) &= z_{22}+\frac{\partial}{\partial z_{22}}\ln p(z_2)\\ &\simeq z_{22}+\frac{\partial}{\partial z_{22}}\ln p_0(z_2)+\epsilon^2\frac{\partial}{\partial z_{22}}\left(\frac{\partial^2}{\partial\epsilon^2}\ln p_\epsilon(z_2)\bigg|_{\epsilon=0}\right)\\ &= z_{22}-(z_{22}-f(z_{21}))\\ &+\epsilon^2\left[f''(z_{21})+2f'^2(z_{21})(z_{22}-f(z_{21}))+2f'(z_{21})\left(-\frac{z_{21}}{\sigma_0^2}\right)\right].\end{aligned} \tag{5.17}$$

Our goal is to get an expansion for the cost, which as we know from (3.5) can be written as

$$J^*(f)=k^2E\left[(z_1-f(z_1))^2\right]+E\left[(f(z_1))^2\right]-E\left[(g^*(z_2))^2\right]. \tag{5.18}$$

Using the expansion for $g^*(z_2)$ from (5.17), we have

$$E\left[(g^*(z_2))^2\right] \simeq E\left[(f(z_{21}))^2\right]$$
$$+ 2\epsilon^2 E\left[f(z_{21})\left(f''(z_{21}) + 2f'^2(z_{21})(z_{22} - f(z_{21})) + 2f'(z_{21})\left(-\frac{z_{21}}{\sigma_0^2}\right)\right)\right], \tag{5.19}$$

where we have neglected the fourth order term in ϵ. Substituting this expansion back in (5.18), we will obtain the following expansion for the cost:

$$J^*(f) = k^2 E\left[(z_1 - f(z_1))^2\right] + E\left[(f(z_1))^2\right] - E\left[(f(z_{21}))^2\right]$$
$$- 2\epsilon^2 E\left[f(z_{21})\left(f''(z_{21}) + 2f'^2(z_{21})(z_{22} - f(z_{21})) + 2f'(z_{21})\left(-\frac{z_{21}}{\sigma_0^2}\right)\right)\right]. \tag{5.20}$$

Note that when the transmission is noiseless, i.e., $\epsilon = 0$ and therefore $z_{21} = z_1$, we have

$$J^*(f) = k^2 E\left[(z_1 - f(z_1))^2\right], \tag{5.21}$$

and $f(z_1) = z_1$ is the obvious unique optimal solution. The above expansion, however, is not exactly in our desired form yet. This is because the third term on the right-hand side, which is an average over z_{21}, still depends on ϵ. We shall now rewrite the expansion in (5.20) by explicitly expressing the expectations based on the corresponding probability densities:

$$J^*(f) = \int_{-\infty}^{+\infty}\left[k^2(t - f(t))^2 + f^2(t)\right]\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{t^2}{2\sigma_0^2}}dt$$
$$- \int_{-\infty}^{+\infty}\left[f^2(t) + 2\epsilon^2\left(f(t)f''(t) - 2f(t)f'(t)\frac{t}{\sigma_0^2}\right)\right]\frac{1}{\sqrt{2\pi(\sigma_0^2+\epsilon^2)}}e^{-\frac{t^2}{2(\sigma_0^2+\epsilon^2)}}dt$$
$$- \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} 4\epsilon^2 f(t)f'^2(t)(\tau - f(t))\frac{1}{\sqrt{2\pi}}e^{-\frac{(\tau-f(t))^2}{2}}\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{t^2}{2\sigma_0^2}}dtd\tau, \tag{5.22}$$

where we have substituted $p(z_2) = p(z_{22}, z_{21}) \simeq p_0(z_2)$ in the third term, since the higher order terms would be multiplied by ϵ^2 and would then be neglected. Now the third term turns out to be zero, because

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} 4\epsilon^2 f(t)f'^2(t)(\tau - f(t))\frac{1}{\sqrt{2\pi}}e^{-\frac{(\tau-f(t))^2}{2}}\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{t^2}{2\sigma_0^2}}dt$$
$$= \int_{-\infty}^{+\infty} 4\epsilon^2 f(t)f'^2(t)\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{t^2}{2\sigma_0^2}}\left(\int_{-\infty}^{+\infty}(\tau - f(t))\frac{1}{\sqrt{2\pi}}e^{-\frac{(\tau-f(t))^2}{2}}d\tau\right)dt = 0. \tag{5.23}$$

At the same time, we can expand the probability density of z_{21} up to the second order in ϵ. It is actually straightforward to obtain

$$\frac{1}{\sqrt{2\pi(\sigma_0^2+\epsilon^2)}}e^{-\frac{t^2}{2(\sigma_0^2+\epsilon^2)}} \simeq \frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{t^2}{2\sigma_0^2}} + \epsilon^2\frac{1}{\sqrt{2\pi}\sigma_0^5}\left(t^2 - \sigma_0^2\right)e^{-\frac{t^2}{2\sigma_0^2}}. \tag{5.24}$$

Substituting (5.23) and the above expansion back in (5.22) and neglecting the higher order terms in ϵ, we can finally get the following expansion for the cost:

$$J^*(f) = \int_{-\infty}^{+\infty} \left[k^2 (t - f(t))^2\right] \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{t^2}{2\sigma_0^2}} dt + \epsilon^2 \int_{-\infty}^{+\infty} \left[4f(t)f'(t)\frac{t}{\sigma_0^2} - 2f(t)f''(t) + f^2(t)\frac{\sigma_0^2 - t^2}{\sigma_0^4}\right] \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{t^2}{2\sigma_0^2}} dt$$

$$:= J_0^* + \epsilon^2 J_1^*. \tag{5.25}$$

The objective is now to obtain the function f, which minimizes the above approximated cost. In the next section, we use a variational approach in order to find a necessary condition for such a function and show how the linear strategies still satisfy this necessary condition.

6. Minimizing the approximated cost. So far, we have obtained an expansion for the cost assuming that the transmission noise intensity is small. We have, in fact, approximated the cost by including only up to the second order term in ϵ. We should now try to minimize this approximated cost in order to find the asymptotically optimal f^*. Obviously, this strategy would be optimal only for a small transmission noise intensity. However, it would still be very helpful for the analysis of the behavior of the optimal strategies when we deviate a little bit from the classical information pattern by introducing a small communication uncertainty.

We now use the Hamiltonian approach in order to find the necessary conditions for the function $f(t)$, which minimizes our approximated cost. For simplicity, denote

$$x_1(t) := f(t), \tag{6.1}$$

$$x_2(t) := \dot{x}_1(t) = f'(t), \tag{6.2}$$

$$u(t) := \dot{x}_2(t) = \ddot{x}_1(t) = f''(t), \tag{6.3}$$

$$p(t) := \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{t^2}{2\sigma_0^2}}. \tag{6.4}$$

The Hamiltonian is then defined as [12]

$$\mathcal{H} = k^2 (t - x_1(t))^2 p(t) + \epsilon^2 \left(4x_1(t)x_2(t)\frac{t}{\sigma_0^2} - 2x_1(t)u(t) + x_1^2(t)\frac{\sigma_0^2 - t^2}{\sigma_0^4}\right) p(t) + \lambda_1(t)x_2(t) + \lambda_2(t)u(t), \tag{6.5}$$

where λ_1 and λ_2 are the Lagrange multipliers that should satisfy

$$\dot{\lambda}_1(t) = -\mathcal{H}_{x_1} = \left(2k^2 (t - x_1(t)) - 4\epsilon^2 x_2(t)\frac{t}{\sigma_0^2} - 2\epsilon^2 x_1(t)\frac{\sigma_0^2 - t^2}{\sigma_0^4} + 2\epsilon^2 u(t)\right) p(t), \tag{6.6}$$

$$\dot{\lambda}_2(t) = -\mathcal{H}_{x_2} = -4\epsilon^2 x_1(t)\frac{t}{\sigma_0^2} p(t) - \lambda_1(t). \tag{6.7}$$

But as we can see, the Hamiltonian is linear in $u(t)$ and we actually have a *singular* optimization problem. The singular surface will be characterized by setting $\mathcal{H}_u$ and its derivatives with respect to t equal to zero, that is,

$$\mathcal{H}_u = -2\epsilon^2 x_1(t)p(t) + \lambda_2(t) = 0, \tag{6.8}$$

and

$$\frac{d}{dt}\mathcal{H}_u = -2\epsilon^2\dot{x}_1(t)p(t) - 2\epsilon^2 x_1(t)\dot{p}(t) + \dot{\lambda}_2(t) = 0. \tag{6.9}$$

Substituting $\dot{p}(t) = -\frac{t}{\sigma_0^2}p(t)$ and also $\dot{\lambda}_2$ from (6.7), we get

$$\frac{d}{dt}\mathcal{H}_u = -2\epsilon^2 x_2(t)p(t) - 2\epsilon^2 x_1(t)\frac{t}{\sigma_0^2}p(t) - \lambda_1(t) = 0. \tag{6.10}$$

Differentiating again and substituting $\dot{\lambda}_1$ from (6.6), we have

$$\frac{d^2}{dt^2}\mathcal{H}_u = -4\epsilon^2 u(t)p(t) + 4\epsilon^2\frac{t}{\sigma_0^2}x_2(t)p(t) - 2k^2\left(t - x_1(t)\right)p(t) = 0. \tag{6.11}$$

Therefore, the corresponding $u(t)$ on the singular surface is

$$u(t) = x_2(t)\frac{t}{\sigma_0^2} - \frac{k^2}{2\epsilon^2}\left(t - x_1(t)\right). \tag{6.12}$$

Note that the first order generalized Legendre–Clebsch condition, which is a necessary condition for $u(t)$ to be minimizing on the singular surface, is also satisfied, namely,

$$\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\mathcal{H}_u\right) \leq 0. \tag{6.13}$$

Therefore, the corresponding $x_1(t)$ and $x_2(t)$, which minimize our approximated cost, should necessarily satisfy the following differential equations:

$$\dot{x}_1(t) = x_2(t), \tag{6.14}$$

$$\dot{x}_2(t) = x_2(t)\frac{t}{\sigma_0^2} - \frac{k^2}{2\epsilon^2}\left(t - x_1(t)\right). \tag{6.15}$$

Since ϵ is assumed to be small, we may assume the following form in order to obtain the solutions for the above differential equations:

$$x_1(t) = a_0(t) + \epsilon^2 a_2(t) + \epsilon^4 a_4(t) + \cdots, \tag{6.16}$$

$$x_2(t) = b_0(t) + \epsilon^2 b_2(t) + \epsilon^4 b_4(t) + \cdots. \tag{6.17}$$

Interestingly enough, by substituting the above x_1 and x_2 back into the differential equations and comparing the coefficients of the terms with the same order in ϵ, we get

$$x_1(t) = \left[1 - \frac{2\epsilon^2}{k^2\sigma_0^2} + \left(\frac{2\epsilon^2}{k^2\sigma_0^2}\right)^2 - \left(\frac{2\epsilon^2}{k^2\sigma_0^2}\right)^3 + \cdots\right]t = \frac{t}{\left(1 + \frac{2\epsilon^2}{k^2\sigma_0^2}\right)}. \tag{6.18}$$

Back to our original notation, we actually have

$$f(z_1) = \frac{z_1}{\left(1 + \frac{2\epsilon^2}{k^2\sigma_0^2}\right)}. \tag{6.19}$$

As we can see, the solution is still linear with a coefficient which is slightly different than the corresponding coefficient for the noiseless transmission case. Remember that

$f(z_1) = z_1$ is the optimal solution when there is no transmission noise, and note that for $\epsilon = 0$ in (6.19) we get exactly the same solution as expected. Given the above function $f(z_1)$, the corresponding $g^*(z_2)$ can easily be obtained using (3.4). Note that it will also be linear because of the Gaussian assumption for the underlying uncertainties.

We could somehow expect the optimal strategies to be linear from the beginning. As we mentioned in section 2, linear strategies were shown to be asymptotically optimal for the Witsenhausen example when the uncertainty on the information available to the first station is small [2]. In this paper, however, we have considered a reformulation of Witsenhausen's problem where the first station sends its information to the second station through a low noise channel. These two scenarios are somewhat similar. Namely, in both scenarios, the second station can determine the information available to the first station fairly accurately. Specifically, in the first scenario, the second station almost knows z_1 because of its small uncertainty, while in the second scenario it can determine z_1 from the information that is transmitted through a low noise channel.

We would also expect the optimal strategies to approach the corresponding strategies for the noiseless transmission case as the value of z_1 and, in some sense, the *signal-to-noise ratio* increases. This does not seem to happen in the solution (6.19). One may justify this by looking at the exponential function in the cost (5.25). This function drives the integrand of the cost to zero exponentially fast for large values of z_1. Therefore, the structure of the cost really does not force the optimal solution to approach $f(z_1) = z_1$ as z_1 increases.

We shall now obtain the corresponding value of the cost. Substituting $f(t)$ from (6.19) back into the cost (5.25), we get

$$
\begin{aligned}
J^*(f) = &\int_{-\infty}^{+\infty} \left[k^2 \left(t - \frac{t}{1+\frac{2\epsilon^2}{k^2\sigma_0^2}} \right)^2 \right] \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{t^2}{2\sigma_0^2}} dt \\
&+ \epsilon^2 \int_{-\infty}^{+\infty} \left[4 \frac{t}{\left(1+\frac{2\epsilon^2}{k^2\sigma_0^2}\right)^2} \frac{t}{\sigma_0^2} + \frac{t^2}{\left(1+\frac{2\epsilon^2}{k^2\sigma_0^2}\right)^2} \frac{\sigma_0^2 - t^2}{\sigma_0^4} \right] \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{t^2}{2\sigma_0^2}} dt \\
&= \frac{1}{\left(1+\frac{2\epsilon^2}{k^2\sigma_0^2}\right)^2} \left(2\epsilon^2 + \frac{4\epsilon^4}{k^2\sigma_0^2} \right) \\
&\simeq 2\epsilon^2 - \frac{4\epsilon^4}{k^2\sigma_0^2}, \qquad (6.20)
\end{aligned}
$$

where we have used

$$
\int_{-\infty}^{+\infty} t^2 \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{t^2}{2\sigma_0^2}} dt = \sigma_0^2, \qquad (6.21)
$$

$$
\int_{-\infty}^{+\infty} t^4 \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{t^2}{2\sigma_0^2}} dt = 3\sigma_0^4. \qquad (6.22)
$$

The optimal cost for the noiseless transmission case is zero. But if we use $f(z_1) = z_1$ when the transmission is noisy, we get the following cost:

$$
J^*(f) = 2\epsilon^2. \qquad (6.23)
$$

In other words, if we fix the strategies to be the optimal strategies for the noiseless transmission case while introducing a small transmission noise, the increase in the cost will be proportional to the transmission noise intensity. However, if we use (6.19), we can indeed improve the cost by the fourth order in ϵ.

One should note from (6.19) and (6.20) that as the value of $k^2\sigma_0^2$ increases, the asymptotically optimal solution approaches $f(z_1) = z_1$, and the change in the cost becomes smaller. In other words, increasing $k^2\sigma_0^2$ has an effect similar to decreasing the communication uncertainty. To explain this, we note from the performance index that increasing k^2 implies a more expensive control action for the first station, which, in turn, results in smaller u_1. This then implies that the information available to the second station is less affected by the action of the first station. At the same time, increasing σ_0^2 implies a higher level of uncertainty on x_0, which, incidentally, is the piece of information that is being transmitted between the stations.

This brings up an example of a very interesting fundamental issue: the notion of *information value* and how it could be different for control and communication purposes. In fact, we know from information theory that a higher level of uncertainty for a piece of information implies a higher level of *entropy* and therefore a more valuable piece of information for transmission. On the other hand, however, a more uncertain piece of information would probably be less valuable for control purposes and would have smaller effect on the control strategies. In other words, a control designer would probably be willing to spend less on installing transmitters on the stations for communicating more uncertain pieces of information. While defining a notion for the value of information for control purposes has been occasionally addressed in the literature for quite a long time, it still remains an open problem. This is mostly because of the fact that the value of information for control purposes would highly depend on how the cost is defined for the control design, and this could be quite different in various applications.

7. Concluding remarks. We analyzed an example of a decentralized stochastic system. This example was a reformulation of the Witsenhausen counterexample where the first station was allowed to send its information to the second station through a noisy channel. The dynamics were linear, all the underlying uncertainties were assumed to be Gaussian, and the cost was quadratic. It was shown that as soon as any uncertainty is introduced in the communication among the stations, the information pattern again becomes nonclassical, which is not partially nested. We then showed how the performance index can be alternatively expressed such that the possible nonconvexity of the cost, with respect to the control strategies, becomes more transparent. Therefore, in general, we will end up with a nonconvex functional optimization problem when we try to obtain the decentralized optimal control algorithms. We then considered two limit cases. Namely, the case where there is no communication uncertainty and the case in which the transmission noise intensity increases to infinity. The former case was shown to be a trivial example of a classical LQG problem, whereas the latter case corresponds to Witsenhausen's counterexample, the optimal solution of which is still unknown.

We then focused on the case where the communication uncertainty was small. We followed an asymptotic approach where we approximated the cost based on its expansion in terms of the small transmission noise intensity. We showed how minimizing the approximated cost can be seen as a singular optimization problem. We then used a variational approach in order to find the necessary conditions for the asymptotically optimal strategies and showed that some reasonable linear strategies

would actually satisfy those conditions. We also provided some intuitive explanations for the behavior of those linear strategies and obtained the corresponding cost.

Note that while we have focused on the reformulated Witsenhausen counterexample, our main result is quite general. In fact, we have shown through an example that communication uncertainties in decentralized systems generally result in nonclassical information patterns, which, in turn, can destroy the convexity of the associated functional optimization problems. Moreover, our approach is indeed a very general approach, which have been applied to various other problems before. More specifically, expanding a cost function in terms of some small parameters is a common practice in variational and perturbation-based approaches. Furthermore, using Hamiltonian approach in order to obtain the necessary conditions for the optimal strategies obviously is not specific to our reformulated Witsenhausen problem. However, finding the exact function (6.19), which is obtained in closed form, satisfies the necessary condition for optimality, and shows how the optimal strategies could change upon introduction of some communication uncertainty, could be very specific to our problem.

All the derivations and the results in this paper show some of the difficulties involved in dealing with decentralized systems as soon as we deviate a little bit from a classical, or at least a partially nested, information pattern. On the other hand, even though we have modeled the communication uncertainty in the simplest possible way, we have tried to emphasize the role of communication uncertainties in generating such information patterns that are very difficult to handle.

Finally, it should be mentioned that even though the optimization problem is generally difficult for this class of systems, in some applications one might be able to exploit the specific structure of the system in order to obtain some reasonably good *suboptimal* strategies, which could yield an acceptable performance.

REFERENCES

[1] H. S. Witsenhausen, *A counterexample in stochastic optimum control*, SIAM J. Control, 6 (1968), pp. 131–147.

[2] D. A. Castanon and N. R. Sandell, Jr., *Signaling and uncertainty: A case study*, in Proceedings of the IEEE Conference on Decision and Control, 1978, pp. 1140–1144.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[4] M. L. Cohen, *The Fisher information and convexity*, IEEE Trans. Inform. Theory, 14 (1968), pp. 591–592.

[5] M. Baglietto, T. Parisini, and R. Zoppoli, *Numerical solutions to the Witsenhausen counterexample by approximating networks*, IEEE Trans. Automat. Control, 46 (2001), pp. 1471–1477.

[6] M. Baglietto, T. Parisini, and R. Zoppoli, *Nonlinear approximations for the solution of team optimal control problems*, in Proceedings of the 36th IEEE Conference on Decision and Control, Vol. 5, 1997, pp. 4592–4594.

[7] M. Baglietto, T. Parisini, and R. Zoppoli, *Neural networks for the solution of information distributed optimal control problems*, in Proceedings of 6th European Symposium on Artificial Neural Networks, ESANN'98, 1998, pp. 79–84.

[8] Y. C. Ho and T. S. Chang, *Another look at the non-classical information structure problem*, IEEE Trans. Automat. Control, 25 (1980), pp. 537–540.

[9] C. H. Papadimitriou and J. Tsitsiklis, *Intractable problems in control theory*, SIAM J. Control Optim., 24 (1986), pp. 639–654.

[10] J. T. Lee, E. Lau, and Y. C. Ho, *The Witsenhausen counter-example: A hierarchical search approach for nonconvex optimization problems*, IEEE Trans. Automat. Control, 46 (2001), pp. 382–397.

[11] J. T. Lee, E. Lau, and Y. C. Ho, *On the Global Optimum of the Witsenhausen Counter-Example*, manuscript.

[12] A. E. Bryson, Jr., and Y. C. Ho, *Applied Optimal Control*, Taylor and Francis, Bristol, PA, 1975.

Appendix B

"A distributed scheme for integrated predictive dynamic channel and power allocation in cellular radio networks"

Kambiz Shoarinejad, Jason L. Speyer, and Gregory J. Pottie,

Proceedings of the IEEE Globecom Conference 2001 and to be publish in the *IEEE Transactions on Wireless Communication (IEEE ToWC).*

Integrated predictive power control and dynamic channel assignment in mobile radio systems*

Kambiz Shoarinejad †
Innovics Wireless Inc.

Jason L. Speyer ‡
UCLA Mechanical and Aerospace Engineering

Gregory J. Pottie §
UCLA Electrical Engineering

Keywords: Cellular Land Mobile Radio, Power Control, Dynamic Channel Allocation, Kalman Filters

Abstract

It is known that dynamic allocation of channels and power in a Frequency/Time Division Multiple Access (FDMA/TDMA) system can improve performance and achieve higher capacity. Various algorithms have been separately proposed for dynamic channel assignment and power control. Moreover, integrated Dynamic Channel and Power Allocation (DCPA) algorithms have already been proposed based on simple power control algorithms. In this paper, we propose a DCPA scheme based on a novel predictive power control algorithm. The Minimum Interference Dynamic Channel Assignment algorithm is employed, while simple Kalman Filters are designed to provide the predicted measurements of both the channel gains and the interference levels, which are then used to update the power levels. Local and global stability of the network are analyzed and extensive computer simulations are carried out to show the improvement in performance, under the dynamics of user arrivals and departures and user mobility. It is shown that call droppings and call blockings are decreased while, on average, fewer channel reassignments per call are required.

I. Introduction

With the ever increasing need for capacity in mobile radio systems, optimal allocation of resources in non-uniform and non-stationary environments has become a great challenge. The fundamental objective is to accommodate as many users as possible, subject to complexity and Quality of Service requirements, on a

*This research was supported in part by the Air Force Office of Scientific Research under Grant Number F49620-00-1-0154
†11500 W. Olympic Blvd., Suite 398, Los Angeles, CA 90064, (kambiz@innovicswireless.com).
‡Box 951597, Los Angeles, CA 90095-1597, (speyer@seas.ucla.edu). Author to whom correspondence should be addressed.
§Box 951594, Los Angeles, CA 90095-1594, (pottie@ee.ucla.edu).

limited available bandwidth by controlling undesired interactions among the users. One major interaction is the co-channel interference that every user generates for all other users, which are sharing the same channel. Various techniques have been developed to mitigate the effects of co-channel interference. Some of these techniques, such as sectorization and beamforming using smart antenna arrays, try to suppress interference, while others such as channel assignment techniques try to avoid strong interferers.

Another well-known technique is to adaptively control the power levels of all the users in the network. The idea is to keep the power level for every user at its minimum required level according to the current channel conditions. This will eliminate unnecessary interference to other users and will also minimize the power consumption for the user. Various power control algorithms have been proposed in the literature [1]-[12].

Our first objective in this paper is to design a distributed predictive power control algorithm. We try to obtain accurate enough models for the slow variations in the channel gains and the interference powers. We then design Kalman filters for every user to obtain the one-step predicted values for both the interference level and the user's channel gain from its intended base station. We try to tune the filters for a typical mobile radio environment and then conjecture and show through simulations that the filters are indeed robust under a broad range of parameters such as user velocities and shadowing correlation distances. The predicted measurements from the Kalman filters are then used in an integrator algorithm to update the power levels.

Another approach to mitigate the co-channel interference effects and increase the capacity is to avoid strong interferers by dynamically assigning the channels to the users. Various centralized and decentralized Dynamic Channel Assignment (DCA) schemes have been proposed in the literature [13]-[16].

It is believed that an aggressive DCA scheme can make an FDMA/TDMA system an *interference-limited* system, where the number of active users is mostly limited by the interference that the users cause on each other. On the other hand, power control schemes are known to be especially effective for interference-limited systems. This has initiated research on integrated distributed Dynamic Channel and Power Allocation (DCPA) schemes [17]-[20]. In [17] a pilot based minimum interference DCA scheme is integrated with a fast fixed-step power control algorithm, while fast fading and user mobility effects are neglected. In [18] three different types of minimum interference DCA algorithms are integrated with a slow integrator power control algorithm. Pedestrian mobility along with a low power update rate are considered and it is again assumed that the fast fading effects are averaged out. In [19] a simulation study has been performed to investigate the joint effects of some simple SIR[1]-based and signal-level-based power control algorithms along with a minimum interference

[1]SIR denotes Signal to Interference plus Noise Ratio throughout this paper.

channel reassignment scheme. Fast fading effects are again neglected and low power update rates are assumed.

Most DCPA schemes, however, only consider simple power control algorithms. Moreover, except for [18]-[19], other results neglect such effects as dynamics of user arrival or departures, user mobility, and base station hand-offs. Our main objective in this paper is to investigate the performance of our predictive power control algorithm when it is integrated with a minimum interference DCA scheme. We set up a system-level simulation platform, similar to the ones presented in [17]-[18], to compare our predictive DCPA scheme with the one that uses a simple integrator power control algorithm with no prediction. Dynamics of user arrivals and departures, user mobility and base station hand-offs are all considered in this study. Slowly varying flat Rayleigh fading effects are also considered in the simulations.

The organization of the paper is as follows. In the next section, we present the system model and review some of the results in power control and dynamic channel assignment. In Section III we elaborate on our predictive power control design. We explain how simple Kalman filters may be designed and implemented in order to obtain the predicted measurements of both the channel gains and the interference powers. We also show that the presented predictive power control algorithm satisfies the sufficient conditions for global stability of the network. In Section IV we describe in detail our simulation models and in Section V we discuss the simulation results and compare the performance of our integrated predictive DCPA algorithm with the corresponding algorithm which uses no prediction. We show that, for a range of traffic loads, the number of blocked calls and dropped calls are decreased under our predictive DCPA scheme. Moreover, on average, fewer number of channel reassignments are required for every call, implying a more stable network. We will provide concluding remarks in the final section.

II. System Model, Dynamic Channel Assignment and Power Control

We consider a cellular system where the area under coverage is divided into cells and each cell has its own base station. All users communicate with their assigned base stations through a single hop. This is in contrast to *ad hoc* wireless networks where there is no fixed infrastructure and multi-hop communication is prevalent.

We focus on a Frequency/Time Division Multiple Access (FDMA/TDMA) system and only consider the co-channel interference among the users, i.e., no adjacent channel interference is assumed. Specifically, we assume a system-wide synchronization to the slot level so that each user will experience interference only from the users which are sharing exactly the same slot on the same carrier frequency. This assumption implies that large enough guard times per slot are assumed. We do not consider any blind slots in the system, that

is, we assume that any slot in a frame can be used as a traffic channel. Blind slots can be avoided either by appropriate structuring of the control channel or by assuming that a call activity detection scheme is employed such that the users can temporarily discontinue their transmission in their active slots. Modifying the frame structure and considering some slots as the blind slots should not have major effects on our performance comparisons.

We focus on the uplink channel, i.e., the channel from mobiles to base stations. Almost all the results and discussions, however, could similarly be stated for the downlink channel. We assume a fixed-power pilot (control) channel on the downlink. As we shall see, this channel facilitates Dynamic Channel Assignment (DCA) and can be used by the mobiles for initial base station assignments and base station hand-offs.

We abstract the system architecture, as far as modulation, coding, etc. are concerned, and consider SIR as the only measure for Quality of Service (QoS) in the system. This is a common practice, even though Bit Error Rate or Frame Error Rate are usually seen as the ultimate performance measures. The reason is that, in general, higher SIR will result in better bit error rate performance and considering SIR as the measure for quality of service provides us with a more convenient platform for power control design.

The received SIR on an assigned uplink channel for user i can now be written as:

$$r_i = \frac{g_{ii}p_i}{\sum_{j=1,j\neq i}^{M} g_{ij}p_j + \eta_i} \tag{1}$$

where p_i is the transmit power for user i, g_{ii} is the channel gain (or attenuation) from user i to its intended base station (in the linear scale), g_{ij} is the channel gain from user j to the intended base station of user i and η_i is the receiver noise intensity at the intended base station of user i. Also M is the total number of users sharing the channel. We now review the minimum interference dynamic channel assignment scheme along with the main approaches for power control.

A. Dynamic Channel Assignment

Under a Dynamic Channel Assignment (DCA) scheme, all base stations have access to all the channels and dynamically assign the channels to the users based on the current traffic conditions. While DCA schemes are clearly more complicated, they usually result in higher capacity.

We adopt a distributed Minimum Interference DCA scheme [15]. In this scheme, the new users will be assigned to the idle channels with minimum local mean interference, in the order they arrive. It was shown in [26] that when a new user is admitted to a power-controlled network, the optimal power level for the new

user can be written as:

$$p_n^* = \frac{I_{n0}}{g_{nn}} \frac{\gamma_n}{1 - \frac{\gamma_n}{\gamma_{max}}}, \tag{2}$$

where γ_n is the SIR threshold that the new user wants to achieve, γ_{max} is the maximum achievable SIR for the new user and I_{n0} is the local mean interference plus noise level at the intended base station of the new user before it is admitted to the network. It is now clear that the minimum interference DCA scheme does indeed result in the minimum transmit power for the new user.

Whenever the local mean SIR for a user drops below a given threshold while the user is transmitting at its maximum power level, a channel reassignment attempt is triggered and, if possible, the user is reassigned to the idle channel, which currently has the minimum local mean interference. Note that this is a distributed scheme, which, in general, is not globally optimal. Remember that any kind of global optimality in the channel assignments can only be achieved through centralized algorithms, which are usually impractical due to the excessive requirements for processing and also communications among the base stations.

Another issue is call management and admission control. As we shall see, a network should be feasible for every user to be able to achieve its desired SIR threshold. If no admission control is employed, a new user could potentially force the network out if its feasibility region and hence result in dropping active calls. Therefore, an admission control mechanism is needed to adjust the trade-off between blocking new calls and dropping active calls. In [21] an admission algorithm was presented for a power controlled system, where the new users would increase their powers only in small steps. It was shown how this scheme could protect the quality of active links when new users arrive. Channel probing techniques were later proposed in [22]-[24], where a new user would try to estimate the maximum SIR level that it can achieve by disturbing the network as little as possible. The user will then be admitted only if its maximum achievable SIR is above its desired threshold. Also a channel partitioning scheme was presented in [25] where a combination of dynamically allocated and fixed assigned channels are incorporated to develop a rapid distributed access algorithm.

We adopt the simpler threshold-based implicit admission control scheme, presented in [18]. In this scheme, a new user with a desired SIR threshold γ_d will be admitted only if there exists an idle channel, on which it can achieve an SIR threshold γ_{new}, which is higher than γ_d by a given protection margin. The value of the protection margin for new users should be selected based on the trade-off between blocking new calls and dropping active calls.

Moreover, a channel reassignment attempt will be triggered for a user if, while transmitting at the maximum power, its local mean SIR drops below a threshold γ_{min}, which is lower than γ_d by another given margin. This

margin is required to avoid excessive number of channel reassignments. The value of this margin should be selected according to the trade-off between quality of service and the average number of channel reassignments per call. Note that for channel reassignment, it is checked whether the user can achieve γ_d on the idle channel which currently has the minimum interference. Since $\gamma_d < \gamma_{new}$, this scheme clearly favors the active users, that are being reassigned, to the new incoming users. If a channel reassignment fails, the user stays on its old channel and the reassignment attempt is repeated every reassignment period (as long as $r < \gamma_{min}$ and $p = p_{max}$) until the user is either successfully reassigned or dropped from the network. Finally, a user will be dropped from the network if its local mean SIR drops and stays below a threshold $\gamma_{drop}(< \gamma_{min})$ for a given duration of time.

B. Power Control

While DCA schemes achieve higher levels of capacity by dynamically distributing the traffic across the channels, power control techniques focus on every channel and try to mitigate the co-channel interference by dynamically adjusting the power levels of the co-channel users at their minimum required levels. Therefore one can reasonably expect that integrating power control with DCA can achieve even higher levels of capacity, even though the capacity gains may not be exactly additive due to some redundancy between the two schemes [18].

A widely studied approach for power control is the *SIR threshold* approach, presented in [4], where the objective is for the SIR of each user in the network to be above a desired threshold, that is:

$$r_i = \frac{g_{ii}p_i}{\sum_{j \neq i} g_{ij}p_j + \eta_i} \geq \gamma_i \tag{3}$$

A necessary and sufficeint condition for the existence of the optimal power levels p_j^*, that satisfy the above set of inequalities, is called *feasibility*. In other words, a network of users is called feasible if every user can achieve its desired SIR. It was shown in [4] that a network is feasible if and only if $\rho(\Gamma(Z - I)) < 1$, where $Z = [z_{ij}] = \left[\frac{g_{ij}}{g_{ii}}\right]$, $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_M)$, $U = [u_i] = [\frac{\gamma_i \eta_i}{g_{ii}}]$, and I is the identity matrix, and ρ denotes the spectral radius of a matrix. Furthermore, under the feasibility condition, the following simple iterative algorithm, which could be implemented in a distributed manner, would converge to the optimal power levels:

$$p_i(n) = \frac{\gamma_i}{g_{ii}} \left(\sum_{j \neq i} g_{ij}p_j(n-1) + \eta_i \right) = \frac{\gamma_i}{g_{ii}} I_i(n) = p_i(n-1)\frac{\gamma_i}{r_i(n)}, \tag{4}$$

where $I_i(n)$ is the total interference plus noise power at the receiver of the intended base station for user i. Therefore, every user only needs a measurement of its own channel gain and its total interference plus

noise in order to update its power level. Note that $I_i(n)$ depends on the power levels of the users during the $(n-1)$-th power update period. Also no extra delays are assumed for processing and propagation. Various generalizations of this algorithms have been presented in the literature. A unified framework along with convergence analysis for some of these algorithms were presented in [5].

In most of these algorithms, it is assumed that all the channel gains stay constant for the duration of the convergence of the algorithm. Therefore, it is implicitly assumed that the fading rate of the channel is much slower than the power update rate. In other words, neither the channel gain variations due to user mobility and fading, nor the measurement errors are taken into account. It was recently shown in [6] that the optimal powers obtained from the SIR balancing approach, under constant gain assumptions, are very close to the the optimal powers that minimize the Rayleigh fading induced outage probability for every link.

Some researchers have tried to analyze and possibly modify the power control algorithms to take into account the channel gain variations and the fading induced measurement errors. In [7] it was shown how the desired SIR for the users may be scaled up to guard against the user mobility effects. In [8] a simulation study was performed to investigate the user mobility effects on slow integrator power control algorithm. In [9] a modification of the distributed SIR balancing algorithm was proposed, which was less sensitive to SIR measurement errors. Also in [10] stochastic measurements were incorporated in the power control algorithm and it was shown that the power levels converge, in the mean square sense, to the optimal power levels. More recently, it was shown in [11] how a simple Kalman Filter may be designed to smooth out the interference measurements. Also in [12] it was mentioned how a minimum-variance power control algorithm may be designed when the channel gain variations are modeled by filtered white noise sequences. Despite all this effort towards analysis and design of power control algorithms in non-stationary environments, most of the results fail to provide a systematic approach.

An alternative approach is to formulate the power control problem as a *decentralized regulator* problem, where the objective is for the SIR of every user to *track* a desired threshold, while the channel gains and the interference levels are changing with time and the SIR measurements can be erroneous. Based on this approach, concepts and design methodologies from control theory have already been used for the analysis of some power control algorithms [27] and design of new algorithms [12][28].

We first note that, in the logarithmic scale, the distributed iterative algorithm in (4) is a simple unity gain integrator algorithm in a closed-loop. Using a bar on the variables to indicate values in dB or dBm, we can

write:

$$\bar{p}_i(n) = \bar{p}_i(n-1) + (\bar{\gamma}_i - \bar{r}_i(n)) \triangleq \bar{p}_i(n-1) + \bar{e}_i(n), \tag{5}$$

where $\bar{p}_i(n)$ is the power level in dBm for user i for the duration of the n-th power update period and $\bar{r}_i(n)$ is the SIR in dB for the same user at the beginning of the n-th power update period:

$$\bar{r}_i(n) = \bar{p}_i(n-1) + \bar{g}_{ii}(n) - \bar{I}_i(n) \tag{6}$$

Moreover, $\bar{I}_i(n)$ is the measured local mean interference plus noise power in dBm, available at the beginning of the n-th power update period:

$$\bar{I}_i(n) = 10\log_{10}\left(\sum_{j\neq i} g_{ij}(n)10^{\frac{\bar{p}_j(n-1)}{10}} + \eta_i(n)\right). \tag{7}$$

The block diagram for a single loop, associated with a single user, is shown in Figure 1. The controller transfer function in this case is:

$$K_i(q^{-1}) = \frac{\bar{P}_i(q^{-1})}{\bar{E}_i(q^{-1})} = \frac{1}{1-q^{-1}}, \tag{8}$$

where q is the shift operator. Therefore, the network can be seen as a set of interconnected local loops. It should be realized that the couplings among the local loops is through the interference function (7), which, in general, is nonlinear. The decentralized regulator formulation of the power control problem can now be presented as the following: "*Design a set of local controllers $K_i(q^{-1})$ such that the SIR for every user, $\bar{r}_i$, tracks a desired threshold $\bar{\gamma}_i$ with a certain performance while the global network remains stable.*"

The local loops in Figure 1 are quite general and can be modified to accommodate different modeling assumptions. For example, extra delay blocks may be inserted in the feedback path to model processing and propagation delays. Moreover, a saturation block may be inserted in the forward path after the controller to model the maximum and minimum power constraints. Also we have implicitly assumed a linear time invariant controller by writing $K_i(q^{-1})$. However, in general, the controller itself can be a nonlinear block, as is the case for *Fixed-Step* power control algorithms. Unfortunately, analysis of stability and convergence of the algorithms, designed via this approach, can be very complicated. Both local and global stability for the network should be analyzed while feasibility of the network and its implications should be addressed.

The global stability of the network implies that all the local loops are stable, but the reverse is not necessarily true. It was shown in [26] that as long as the network stays feasible, i.e., the channel gain variations do not force the network out of its feasibility region, a sufficient condition for global stability of the network is:

$$\left|\left|G_i(q^{-1})\right|\right|_{\ell_\infty - induced} \leq 1, \tag{9}$$

where $G_i(q^{-1})$ is the transfer function from the interference $\bar{I}_i(n)$ to power $\bar{p}_i(n-1)$ and the $\ell_\infty - induced$ norm for the single-input-single-output system can be obtained as:

$$\left\|G_i(q^{-1})\right\|_{\ell_\infty - induced} = \|g_i\|_1 = \sum_{k=0}^{\infty} |g_i(k)|, \tag{10}$$

where g_i denotes the impulse response associated with the transfer function G_i.

Hence if the local loops are stable, and if the feasibility condition is not violated and (9) is satisfied for all local loops, then the network will be globally stable in the sense that the deviations of the power levels from their corresponding optimal values will always remain bounded. It was also shown in [12] that if the channel gains are constant and the network is feasible (i.e., a fixed optimal power vector $\bar{P}^*$ exists) and if the interference function (7) is linearized around $\bar{P}^*$, then all small deviations of the power levels in the network from their corresponding optimal values will asymptotically converge to zero if:

$$\left\|G_i(q^{-1})\right\|_{\ell_2 - induced} = \sup_{\omega} \left|G_i\left(e^{j\omega}\right)\right| \leq 1. \tag{11}$$

The above condition is indeed a sufficient condition for global stability of the linearized network in the $\ell_2 - induced$ norm sense, while (9) gives a sufficient condition for global stability in the $\ell_\infty - induced$ norm sense without any linearization or any constant gain assumption.

III. Predictive Power Control

Our objective in this section is to show how simple models for the variations in the channel gains and the interference levels may be used in designing simple Kalman filters, that provide predicted measurements for both the channel gains and the interference levels while they mitigate the effects of the fast fading induced measurement errors.

We are assuming that the received SIR measurement or the power command are sent back to the transmitter. In other words, we are considering *information-feedback* closed-loop power control algorithms. Due to the limitations on the control bandwidth and on the processing time, information-feedback algorithms usually run at slower power update rates. Therefore, similar to DCA algorithms, they operate on the local mean values, which are obtained through some sort of averaging of the measurements over some relatively long periods.

A. Models for Variations in Channel Gains and Interference Levels

The variations in the channel gains can be characterized by the slowly changing shadow fading and the fast multipath fading on top of the distance loss. We consider log-normal shadowing whose spatial (or temporal) correlation is represented with a simple first-order Markov model presented in [30].

The channel gain from every user i to its intended base station, in the logarithmic scale, is therefore modeled as:

$$\bar{g}_{ii}(n) = \bar{g}_{ii}^0 + \delta\bar{g}_{ii}(n) \tag{12}$$

$$\delta\bar{g}_{ii}(n) = a\delta\bar{g}_{ii}(n-1) + w_g(n-1), \tag{13}$$

where $\bar{g}_{ii}^0$ is a constant bias and w_g is a zero mean white Gaussian noise sequence. The constant bias accounts for the antenna gains and the distance loss in the filter. The parameter a is obtained as:

$$a = e^{-\frac{vT}{X_s}}, \tag{14}$$

where v is the user velocity and T is the update period. Note that vT is the distance that the user moves during one update period. Moreover X_s is called the shadowing *correlation distance.* It is the distance at which the normalized correlation decreases to e^{-1}. To see this, note that the autocorrelation function for $\delta\bar{g}$ can be obtained as:

$$R_{\delta\bar{g}}(m) \triangleq \mathrm{E}\left[\delta\bar{g}(m+n)\delta\bar{g}(n)\right] = \frac{\sigma_{w_g}^2}{1-a^2}a^{|m|} = \sigma_s^2 a^{|m|}, \tag{15}$$

where σ_{w_g} denotes the standard deviation of the noise sequence w_g. Note that given the standard deviation for shadowing σ_s and the value for a, the standard deviation for the driving white noise sequence can be obtained.

In order to design distributed algorithms, we need to decouple the local loops in the network. For this purpose, the interference plus noise should be modeled independently for every user. One approach is to treat interference plus noise simply as a bounded disturbance for every user and design the power control algorithm based on the worst case considerations. However, we decide to model the interference plus noise, similar to the channel gains, by white noise driven first-order Markov variations on top of a constant bias. That is:

$$\bar{I}_i(n) = \bar{I}_i^0 + \delta\bar{I}_i(n) \tag{16}$$

$$\delta\bar{I}_i(n) = a\delta\bar{I}_i(n-1) + w_I(n-1), \tag{17}$$

where w_I is a zero-mean white Gaussian noise sequence independent of w_g, but with the same variance. While this model may not exactly capture the slow variations in the interference in a power-controlled system, it can still be reasonable when such slow fluctuations in the interference levels are dominated by shadow fading. Note that, putting aside the changes in the transmit power levels, due to power control, the fluctuations in the channel gains and interefernce levels basically result from the same physical phenomenon. We therefore use

this model in a Kalman Filter to obtain the one-step predicted measurements of the local mean interference values.

Note that one shall use receiver diversity techniques to combat fast fading, since power control algorithms, in general, cannot track very fast channel variations. While we will evaluate the simulated performance of our algorithm with higher power update rates, we decide to select the power update period such that the fast multipath fluctuations are averaged out while the slower shadowing fluctuations are being tracked. It was shown in [31] that, under the flat Rayleigh fading assumption, when a first order low-pass filter or simply a moving average filter is used to obtain the local mean values of the measurements, the averaging error in dB will have a Gaussian distribution, whose mean can be made zero by appropriate choice of the filter DC gain and whose standard deviation depends on the shadow fading standard deviation σ_s, the ratio of the shadowing correlation distance to the carrier wavelength X_s/λ, and the normalized measurement time $f_m T$, where $f_m = v/\lambda$ is the maximum Doppler frequency.

It is now clear that the model parameters not only depend on the environment through the values of the shadowing standard deviation and the shadowing correlation distance, but also depend on the user velocity. While one can think of implementing individual *adaptive* Kalman filters for each user, where the model parameters are continuously updated based on the available information about the user velocities, we choose to consider a fixed model to design and implement the same filters for all the users in the network. There are two main reasons for this. One is that for a rather broad range of user velocities, the values for a and σ_{w_g}, and as shown in [31], the averaging error variance only slightly change and we believe that the Kalman filters will be robust to such changes. The other reason is that while some techniques have been already proposed for user velocity estimation in mobile environments (refer to [32] and the references therein), most of them fail to provide accurate estimates in real time.

B. Kalman Filter Design

Using a set of available measurements, corrupted with Gaussian noise, a Kalman filter recursively obtains the minimum mean squared error estimates of a set of variables that are varying according to a given dynamic model. Kalman filters have proved to be strong estimation tools in a very wide range of applications [33]. As examples of applications in communication systems, Kalman filters have been used for channel equalization [34], interference estimation for call admission in CDMA networks [35] and for power control in packet-switched broadband TDMA networks [9].

We propose a predictive power control algorithm, where two Kalman filters are employed to provide the one-step predicted estimates of both the channel gains and the interference levels for every user, which are then used in an integrator algorithm to update the power levels. Using (12) and (13) for the channel gains, we can write:

$$\bar{g}_{ii}(n) = a\bar{g}_{ii}(n-1) + (1-a)\bar{g}_{ii}^{0} + w_g(n-1). \tag{18}$$

Similarly, using (16) and (17) for the interference levels, we can write:

$$\bar{I}_i(n) = a\bar{I}_i(n-1) + (1-a)\bar{I}_i^{0} + w_I(n-1). \tag{19}$$

The idea is to design two simple Kalman filters that use the erroneous local mean measurements, available to every user, to estimate the constant biases in the models and provide the one-step predicted estimates of the channel gains and the interference levels. As mentioned, the same models are used for all the mobiles in the network. Hence we eliminate the indices i and ii for a simpler notation.

It is now appropriate to represent both models in the state-space form. Define $x_{g1}(n) \triangleq \bar{g}(n)$, $x_{g2}(n) \triangleq \bar{g}^0$, $x_{I1}(n) \triangleq \bar{I}(n)$, and $x_{I2}(n) \triangleq \bar{I}^0$. The state-space models for every user can then be obtained as:

$$x_g(n) = A_f x_g(n-1) + w_g(n-1), \tag{20}$$

$$y_g(n) = H_f x_g(n) + v_g(n), \tag{21}$$

$$x_I(n) = A_f x_I(n-1) + w_I(n-1), \tag{22}$$

$$y_I(n) = H_f x_I(n) + v_I(n) \tag{23}$$

where:

$$x_g \triangleq \begin{bmatrix} x_{g1} \\ x_{g2} \end{bmatrix}, \quad w_g \triangleq \begin{bmatrix} \bar{w}_g \\ w_{g0} \end{bmatrix}, \quad w_I \triangleq \begin{bmatrix} \bar{w}_I \\ w_{I0} \end{bmatrix}, \tag{24}$$

$$A_f \triangleq \begin{bmatrix} a & 1-a \\ 0 & 1 \end{bmatrix}, \quad H_f \triangleq \begin{bmatrix} 1 & 0 \end{bmatrix}, \tag{25}$$

where w_{g0} and w_{I0} are two mutually independent *fictitious* zero mean white Gaussian noise sequences , which are also independent from w_g and w_I. They are required to make the filters more robust to the uncertainties in the models. Moreover, v_g and v_I are mutually independent zero mean white Gaussian noise sequences, which are assumed to be independent from all other noise sequences in the model and are used to model the fast fading induced averaging errors and other possible uncertainties in the local mean measurements. Remember that all the variables are expressed in a logarithmic scale.

Now starting from initial estimates $\hat{x}_g(0)^-$ and $\hat{x}_I(0)^-$, the *measurement update* equations for the filters are expressed as:

$$\hat{x}_g(n)^+ = \hat{x}_g(n)^- + L_g(n)\left(y_g(n) - H_f\hat{x}_g(n)^-\right) \quad (26)$$

$$\hat{x}_I(n)^+ = \hat{x}_I(n)^- + L_I(n)\left(y_I(n) - H_f\hat{x}_I(n)^-\right) \quad (27)$$

where $\hat{x}_g(n)^-$ and $\hat{x}_I(n)^-$ respectively denote the *propagated* (*a priori*) estimates of the channel gain and the interference level at the end of the $(n-1)$-th power update period. Hence, at time n (i.e., the beginning of the n-th power update period), the current local mean measurements $y_g(n)$ and $y_I(n)$ are incorporated to obtain the *updated* (*a posteriori*) estimates $\hat{x}_g(n)^+$ and $\hat{x}_I(n)^+$. The two-dimensional filter gain vectors L_g and L_I are obtained as:

$$L_g(n) = P_g(n)^- H_f^T \left(H_f P_g(n)^- H_f^T + V_g\right)^{-1}, \quad (28)$$

$$L_I(n) = P_I(n)^- H_f^T \left(H_f P_I(n)^- H_f^T + V_I\right)^{-1}, \quad (29)$$

where V_g and V_I are the measurement noise covariances and $P_g(n)^-$ and $P_I(n)^-$ are the *propagated* estimation error covariance matrices. Note that we only have scalar measurements and no matrix inversion is involved. At time n, the covariance matrices are updated as:

$$P_g(n)^+ = P_g(n)^- - L_g(n)H_f P_g(n)^- \quad (30)$$

$$P_I(n)^+ = P_I(n)^- - L_I(n)H_f P_I(n)^-. \quad (31)$$

Now the one-step predicted estimates for the channel gain and the interference level are obtained by propagating the estimates to the next power update period:

$$\hat{x}_g(n+1)^- = A_f\hat{x}_g(n)^+ \quad (32)$$

$$\hat{x}_I(n+1)^- = A_f\hat{x}_I(n)^+, \quad (33)$$

and the covariance matrices are propagated as:

$$P_g(n+1)^- = A_f P_g(n)^+ A_f^T + W_g, \quad (34)$$

$$P_I(n+1)^- = A_f P_I(n)^+ A_f^T + W_I, \quad (35)$$

where W_g and W_I are two-dimensional diagonal covariance matrices for the driving noise sequences in (20) and (22), respectively.

Incorporating the one-step predicted estimates in the integrator algorithm (5), the updated power level for the duration of the n-th power update period can be obtained as:

$$\bar{p}(n) = \bar{p}(n-1) + \left(\bar{\gamma} - \hat{\bar{r}}(n+1)^{-}\right), \tag{36}$$

where:

$$\begin{aligned} \hat{\bar{r}}(n+1)^{-} &= \bar{p}(n-1) + \hat{x}_{g1}(n+1)^{-} - \hat{x}_{I1}(n+1)^{-} \\ &= \bar{p}(n-1) + \hat{\bar{g}}(n+1)^{-} - \hat{\bar{I}}(n+1)^{-}. \end{aligned} \tag{37}$$

When a call is assigned (or reassigned) to an idle channel, its Kalman filter estimates are initialized (or reset) as $\hat{x}_{g1}(0)^{-} = \hat{x}_{g2}(0)^{-} = \bar{g}(0)$ and $\hat{x}_{I1}(0)^{-} = \hat{x}_{I2}(0)^{-} = \bar{I}(0)$, where $\bar{g}(0)$ and $\bar{I}(0)$ are the local mean channel gain and interference values available at the time of channel assignment. Also the error covariance matrices are initialized as $P_g(0)^{-} = P_I(0)^{-} = \mathrm{diag}(\sigma_s^2, \sigma_s^2)$ where σ_s is the shadow fading standard deviation (set to 8 dB in our simulations).

We pick the model parameter a according to (14) and by considering the maximum user velocities that we expect in our mobile environment. This makes the filter assume the least correlation among the local mean values in two consecutive power update periods and therefore rely more on the measurements. As we shall explain in our simulation details, we assume the power levels to be updated every 100 msec. Also we consider the shadowing correlation distance to be about 40m and the maximum user velocity to be 80 km/hr. Using (14), we then pick $a = 0.95$. Using this value for a and $\sigma_s = 8$ dB and (15), we get $\sigma_{w_g}^2 = \sigma_{w_I}^2 = 1.56$. We choose to set $\sigma_{w_g}^2 = \sigma_{w_I}^2 = 2.0$ in the filter, again to deal with uncertainties in the models. The variances for the fictitious driving noise sequences w_{g0} and w_{I0} are also set to 2.0 dB^2. Also the standard deviations for the local mean measurement errors are both set to 3.0 dB, i.e., $V_g = V_I = 9.0$.

One should observe that the error covariance matrices and the filter gains are independent of the actual measurements. This can be seen from the filter equations (28)-(35). Therefore, the filter gains L_g and L_I can, in fact, be calculated and saved *a priori*. This can result in a significant reduction in the filter processing time.

Also note that when the filter reaches the steady-state on a specific channel, the steady-state filter gain vectors are equal to:

$$L_g = L_I = PH_f^T \left(H_f P H_f^T + V\right)^{-1}, \tag{38}$$

where $V_g = V_I = V$ and P is the positive-definite solution to the following *discrete Riccati* equation:

$$P = A_f P A_f^T - A_f P H_f^T (H_f P H^T + V)^{-1} H_f P A_f^T + W, \tag{39}$$

where $W_g = W_I = W$. Using our selected values, we get:

$$L_g = L_I = L = \begin{bmatrix} 0.37990 & 0.37121 \end{bmatrix}^T. \tag{40}$$

C. Global Stability of the Network

When the Kalman filters are employed, the block diagram for a single loop can be depicted as in Figure 2. We now show that, in the steady-state, the Kalman filters and therefore the local loops are stable[2]. Moreover, the sufficient conditions for global stability are satisfied.

Given the filter gains in (40), it is straightforward to obtain the steady-state transfer functions for the Kalman filters:

$$\frac{\hat{\bar{g}}(n+1)^-}{\bar{g}(n)} = \frac{\hat{\bar{I}}(n+1)^-}{\bar{I}(n)} = \frac{q\,(0.37947q - 0.36091)}{q^2 - 1.57053q + 0.58909}. \tag{41}$$

The poles of the Kalman filters (i.e., the poles of the above transfer function or equivalently the eigenvalues of $A_f - A_f L H_f$) are located inside the unit circle at:

$$s_{f1} = 0.61928, \qquad s_{f2} = 0.95125. \tag{42}$$

It is now clear that all the local loops are stable, i.e., the poles for all the closed-loop transfer functions, associated with a single loop, are inside the unit circle. Processing and propagation delays (i.e., extra delay blocks in the feedback path) could result in instability of the local loops and therefore instability of the whole network. However, even though some delay compensation schemes have been proposed in [12], information-feedback power control algorithms, as mentioned before, usually run on lower power update rates and processing and propagation delays are usually much lower than a power update period.

As we mentioned, stability of the local loops is necessary but not sufficient for global stability of the network. However, the network will indeed be globally stable in the $\ell_\infty - \textit{induced}$ norm sense, if the transfer function from the interference $\bar{I}(n)$ to the power $\bar{p}(n-1)$, satisfies the norm condition (9).

Using (41) and from Figure 2, it is straightforward to obtain:

$$G(q) = \frac{\bar{p}(n-1)}{\bar{I}(n)} = \frac{0.37947q - 0.36091}{q^2 - 1.57053q + 0.58909}, \tag{43}$$

and hence we get:

$$||G(q)||_{\ell_2 - \textit{induced}} \simeq ||G(q)||_{\ell_\infty - \textit{induced}} = 1.0. \tag{44}$$

[2]Under the technical conditions of stabilizability and detectability, the steady-state Kalman filters are always known to be stable [33]

Therefore $G(q)$ satisfies both (9) and (11). From (9), we conclude that, as long as the network is in its feasible region, the deviations of the power levels of all the users in the network from their corresponding optimal values will always remain bounded. Moreover, from (11), we conclude that if the power levels only slightly deviate from their optimal values, while the channel gains remain constant, they will asymptotically converge back to their optimal values. This proves the global stability of the network, on every channel, both in ℓ_∞ sense and in ℓ_2 sense (with a linearized interference function), when the Kalman filters are at their steady-state.

When multiple channels are considered and the power control algorithm is integrated with a DCA scheme, the global stability analysis for the network becomes extremely complicated. Average number of channel reassignments per call can be considered as a measure, which can somehow show the level of stability for the network. We show through computer simulations that the average number of channel reassignments per call will be significantly reduced when the Kalman filters are employed in the power control algorithm.

IV. Simulation Model

While the previous theoretical analysis helps in justifying the use of Kalman filters in power control algorithms to deal better with the variations in the channel gains and the interference levels and also the errors in the local mean measurements, a simulation study is essential to analyze the overall performance when such a predictive power control algorithm is integrated with a DCA scheme in a relatively realistic mobile environment. We therefore set up a system-level simulation environment, similar to the ones presented in [17]-[18] but on a smaller scale, in order to analyze the overall performance of the network, when our predictive power control algorithm is integrated with a distributed minimum interference DCA scheme. User arrivals and departures and user mobility are all considered. In this section, we explain the details of our simulation platform and in the next section, we analyze the results.

The simulations run on the frame level, and hence only power and interference levels are simulated and no modulation and coding are considered in the simulations. While we do not restrict ourselves to any specific standard, we have tried to stay close to the *Global System for Mobile Communications* (GSM) standard.

A 3x3 square grid of cells is assumed. The base stations are located on the cell centers and are separated by 800m. To avoid edge effects, a ring simulation structure is assumed, i.e., the statistics are only gathered from the central cell. This is somewhat simpler than a toroidal simulation structure and is shown to provide more optimistic but comparable results [36]. The other reason for our results to be somewhat optimistic is that only nine cells are simulated, and therefore lower interference levels are generated. However, our simulation results

clearly serve our purpose of comparing our predictive DCPA scheme with the one that uses no prediction. Omni-directional antennas with two branch selection diversity is assumed for the base stations.

Every channel is characterized by a pair (m, n) where m denotes the carrier frequency and n is the index for the time slot. We consider two carrier frequencies and eight slots per carrier. As mentioned before, no blind slots are considered. Hence, there are 16 available channels, all of which can potentially be used as traffic channels.

Every frame is 4.0 msec, consisting of 8 slots, each with a duration of 0.5 msec. It is assumed that the signal and interference power measurements for every user are available in every frame at the end of the user's corresponding slot. Various events might then happen every multiple number of frames.

The channel gain for every link is normalized with respect to the base station and mobile antenna gains and is characterized by three components: distance loss, slow or shadow fading and fast fading. The distance loss is assumed to be inversely proportional to d^{α}, where α is set to 4.0. For shadowing a log-normal pattern is generated *a priori*. Therefore the shadowing values only depend on the user's location. The resolution of the shadowing grid is set to be equal to the shadowing correlation distance X_s, which is assumed to be 40m. The shadowing for every user is then obtained by a normalized bilinear interpolation of the four closest points on the shadowing grid. A slowly varying flat Rayleigh fading is also assumed. This implies that no line-of-sight exists and the delay spread is small compared to the symbol duration or the inverse channel bandwidth and thus only a single path with a Rayleigh distributed amplitude (and hence exponentially distributed power) can be distinguished. In fact, the Rayleigh fading component is assumed to be constant for the whole duration of a single slot (0.5 msec). Time correlation for Rayleigh fading is often represented using the Jake's model [29], where it is expressed in terms of a zero order Bessel function of the first kind, which results in a non-rational spectrum. We use a first-order approximation by passing a white complex Gaussian noise through a first order filter and obtaining the squared magnitude of the output Gaussian process. The time constant of the filter, for every user, is obtained by setting its 3 dB cut-off frequency equal to $f_m/4$ where $f_m = v/\lambda$ is the maximum Doppler frequency for the user [13].

New calls are generated based on a Poisson process with a given arrival rate λ_a. Each call is assigned an exponentially distributed holding time with a given average value T_h. The average *Erlang* load per cell is then obtained as $E_c = \lambda_a T_h / N_c$, where $N_c = 9$ is the total number of cells. The Erlang load per cell effectively determines the average number of users that could be active in every cell at any instant of time. We have considered various combinations of values for λ_a and T_h to simulate the network under different traffic load

conditions.

The new users are uniformly distributed in the area. The mobility of the user i is modeled with a constant but random speed v_i and the angle θ_i between the velocity vector and the horizontal axis ($-\pi \leq \theta_i < \pi$). The speed for every new user is selected randomly from a triangular distribution in the range 0-80 km/h. This is preferred over a uniform distribution, as it results in a smaller variance for the velocity distribution among different users. The initial direction θ is uniformly picked. Then every 10 sec, a new direction is selected from a triangular distribution with the old direction as its mean. This is again preferred over a uniform distribution or a two dimensional random walk, since it makes small angle turns more probable that large ones. The motion trajectory for a sample user is shown in Figure 3.

The desired SIR threshold for all users in the network is set to $\bar{\gamma}_d = 12$ dB, while the minimum tolerable SIR is considered to be $\bar{\gamma}_{min} = 10$ dB. Both margins for new user admissions and user droppings are set to 2 dB. Therefore new users will be admitted only if they can achieve $\bar{\gamma}_{new} = 14$ dB on the idle channel with the minimum local mean interference. Moreover, a user will be dropped from the network if its SIR drops below $\bar{\gamma}_{drop} = 8$ dB and stays below for 4.0 consecutive seconds. Note that these margins should have been expressed as percentages of $\bar{\gamma}_d$ and $\bar{\gamma}_{min}$ for every user, if the users were to have different quality of service requirements and thus different SIR thresholds.

When a new user arrives into the network, it first starts scanning the downlink control channel from all neighboring base stations and measures all the local mean channel gains. It is assumed that this process take about 0.8 sec (200 frames), which is called the initial call set-up time. The new user then sends its request for a channel to the base station which has the strongest signal. If this base station does not have any idle channels, the user will try the second best base station. This procedure is called *Direct Retry* and will be repeated for a given number of base stations (set to 3 in our simulations) before the user is blocked. When there are idle channels available, the base station checks whether the user can achieve γ_{new} on the idle channel with the minimum local mean interference. If so, the user will be admitted and will be assigned to the idle channel with the minimum interference. Otherwise, the user will be blocked.

We should note that no *macro diversity* is considered, i.e., any user will only communicate with a single base station at any instant of time. Moreover, base station assignment is considered to be separate from power control, i.e, the power levels are obtained assuming that the users are already assigned to their corresponding base stations. Joint base station assignment and power control has already been proposed in the literature [37].

A minimum interference DCA scheme is employed. The local mean channel gain and interference values for possible channel reassignments are obtained by simple averaging of the available measurements over 50 consecutive frames for every user.

Finally, a base station hand-off attempt will be triggered if the local mean channel gain from a neighboring base station exceeds the corresponding value from the current base station by a selected hand-off margin of 4 dB. If the hand-off attempt fails, the user will stay with its current base station. Note that the users are assumed to be continuously monitoring the downlink control channels of all neighboring base stations.

Two power control algorithms are simulated. Namely, the simple integrator algorithm in (5) and (6) and the predictive algorithm in (36) and (37) are compared. Note that while the propagation simulation models are tailored to individual users, according to their different trajectories and speeds, the same Kalman filter models and parameters are employed for all the users in the network.

After a new user i is admitted, it sets its initial power at:

$$p_i(0) = \frac{\gamma_d I_i(0)^-}{g_{ii}(0)^-}, \tag{45}$$

where $I_i(0)^-$ and $g_{ii}(0)^-$, respectively, denote the local mean channel gain and the interference plus noise level, which are available at the time of user admission. Note that this is somehow an optimistic choice, since a new user sets its initial power as though other users will not increase their transmit powers.

For most of the simulations, the power update rate is assumed to be the same for all users and is set to 100 msec, that is, every user updates its power level every 25 frames. The idea is to have fast multipath fluctuations averaged out while slower variations are being tracked. In all simulations, a maximum transmit power constraint at 30 dBm is imposed on all users in the network, while the receiver noise floor is set to -120 dBm.

It should be mentioned that since the users arrive at arbitrary instants of time according to a Poisson arrival process, the power updates are in fact performed *asynchronously*, even though all the users have the same power update rates. While most results in power control assume synchronous power updates among the users, asynchronous power control algorithms have been addressed in the literature [5]. To have synchronous power updates, one could simply force the users to arrive at instants of time, which are multiples of a common power update period.

V. Performance Analysis

In this section we present and analyze our simulation results and show how the predictive DCPA scheme can improve the overall performance of the network.

For any given traffic load, we run the simulations multiple times with different random generator seeds and every run continues until enough number of calls are dropped. The statistics are then gathered from the central cell.

Figures 4 and 5 show the call blocking and the call dropping responses of the network under the two DCPA schemes. It can be seen that at 7.0 Erlang/Cell, the predictive DCPA scheme achieves about 0.5% lower blocking rate and about 0.03% lower dropping rate. Moreover the improvement in performance increases as the traffic load increases. Remember that there is always a trade-off between blocking new calls and dropping active calls.

Our predictive DCPA scheme also results in better target SIR tracking. We obtain an estimate for the SIR error standard deviation and also estimates for the SIR cumulative distribution functions by looking at the local mean SIR values of all the users in the network at various random instants of time (after enough call attempts have been made and the network has reached some kind of steady state) during every run of the simulation. Figure 6 shows the standard deviation for the error in the local mean SIR for a range of traffic loads. It can be seen that the predictive scheme decreases the SIR error standard deviation by about 0.3 dB at 7.0 Erlang/Cell, while the improvement is about 0.7 dB at 10.0 Erlang/Cell. Furthermore, Figure 7 shows the cumulative distribution for the local mean SIR values in the network under 8.0 Erlang/Cell and 10.0 Erlang/Cell. These figures show how the local mean SIR values for different users are spread around the target SIR value $\gamma_d = 12$ dB. It can be seen that the predictive DCPA scheme results in the local mean SIR values, which are less spread around the target SIR. The improvement is again more noticeable in higher traffic load. In fact, Figure 8 shows how the local mean SIR cumulative distribution function changes with the traffic load under both schemes.

One measure that somehow shows the level of stability of the network is the average number of channel reassignments per call. Figure 9 shows this number for a range of traffic loads under both DCPA schemes. As one would expect, fewer channel reassignments per call are, on average, required in the predictive DCPA scheme. One reason for this is that, as shown before, the predictive scheme does indeed result in better target SIR tracking and smoother local mean SIR behavior.

We also compare the transmit power distribution of the users in the network under the two DCPA schemes.

Figure 10 shows an estimate of the cumulative distribution function for the transmit powers of the users in the network at the load of 8.0 Erlang/Cell. It can be seen that the two schemes perform quite similarly, as far as transmit powers are concerned. In fact, both algorithms result in considerable power saving, when compared with a network where all the power levels are fixed at their maximum levels. For example, at a relatively high load of 8.0 Erlang/Cell, about 50% of the users under both DCPA schemes are transmitting at 0 dBm or lower power levels. It should however be mentioned that our predictive DCPA algorithm seems to result in slightly higher power levels in the network. While one may see this as a small cost for better SIR tracking and better call blocking and dropping responses, it should also be noted that our predictive DCPA scheme does indeed result in higher capacity which in turn implies more active users at any instant in time. This higher traffic explains the higher average transmit power for the users. In fact, Figure 11 shows how the power cumulative distribution functions might change as the traffic load on the network changes under the two DCPA schemes.

Finally, one might argue that our power update rate is too low for the average speeds considered in our simulations. In order to further evaluate the performance of our predictive algorithm, as comapred to standard fast power control schemes, we also simulated the DCPA scheme with standard fixed-step power control algorithm where, depending on the daviation of the received SIR from its target value, the power of each user is incremented or decremented by a fixed 1 dB step every single frame (i.e., once per 4 msec). We then ran the same simulations with our integrated predictive DCPA scheme where the power of each user is updated every 5th frame (i.e., once every 20 msec). Tables 1 and 2 show the call dropping and call blocking probabilities for the two scenarios under two sample traffic loads. It can be seen that the results are similar with our predictive algorithm still performing slightly better. Note however that while some additional computational cost is associated with our algorithm, the update rate for our algorithm is taken to be five times slower than the standard fixed-step algorithm. We do however believe that clarification of the exact trade-off between the extra computation and the lower update rate would require further analysis using simulations and, possibly profiling the code on specific processors.

VI. Conclusion

A predictive Dynamic Channel and Power Allocation scheme was presented in this paper. Simple Kalman filters were designed to obtain the predicted estimates of the local mean channel gains and the local mean interference plus noise levels. These predicted estimates were then incorporated in an integrator algorithm to update the power levels of all the users in the network. It was shown how generic models may be used and

filter parameters may be selected to design the same robust filter for all users. Local stability of the network was analyzed. Moreover it was shown that the sufficient conditions for global stability of the network were satisfied when the Kalman filters were employed in the power control algorithm. The global stability results imply that, as long as the network stays feasible, the deviations of the power levels from their corresponding optimal values will always remain bounded, while the small deviations will always converge back to zero.

This predictive power control algorithm was integrated with a Minimum Interference Dynamic Channel Assignment scheme in an FDMA/TDMA mobile radio system. A system-level simulation environment was then developed. User arrival and departures and user mobility along with flat Rayleigh fading effects were all included in the simulations. It was shown that the predictive DCPA scheme results in better call dropping and call blocking responses and also better target SIR tracking performance for the network. Moreover, on average, fewer channel reassignments per call are required under the predictive DCPA scheme. We believe that these improvements are obtained mainly because the predictive algorithm takes into account at least the slow variations of the channel gains. Also by dealing with uncertainties in the measurements, it effectively mitigates the fading induced local mean measurement errors. It was shown however that the predictive DCPA scheme results in slightly higher power levels for the users in the network.

As for future research, one may try to design adaptive algorithms where the parameters of the algorithm and even the power update rates are adaptively adjusted for individual users, according to such information as user velocities, etc. Also the standard integrator algorithm may not be the best power control algorithm. Even though constraints on complexity and computational effort are always present, other simple algorithms may still be designed that could result in better SIR tracking, better allocation of resources and finally higher levels of capacity in highly non-uniform and non-stationary environments. Finally, analyzing the behavior of any prediction filter, both in terms of convergence and performance, under bursty interference conditions can constitute another interesting line of future research.

References

[1] Zander, J., "Performance of optimum transmitter power control in cellular radio systems," *IEEE Trans. on Vehicular Technology*, Vol. 41, No. 1, pp. 57-62, February 1992.

[2] Zander, J., "Distributed co-channel interference control in cellular radio systems," *IEEE Trans. on Vehicular Technology*, Vol. 41, No. 3, pp. 305-311, August 1992.

[3] Grandhi, S. A., Vijayan, R. and Goodman, D. J., "Distributed power control in cellular radio systems," *IEEE Trans. on Comm.*, Vol. 42, No. 2, pp. 226-8, February-April 1994.

[4] Foschini, G. J. and Miljanic, Z., "A simple distributed autonomous power control algorithm and its convergence," *IEEE*

Trans. on Vehicular Technology, Vol. 42, No. 4, pp. 641-646, November 1993.

[5] Yates, R. D., "A framework for uplink power control in cellular radio systems," *IEEE J. Select. Areas in Comm.*, Vol. 13, No. 7, pp. 1341-7, September 1995.

[6] Kandukuri, S., and Boyd, S., "Optimal power control in interference limited fading wireless channels with outage probability specifications," *IEEE J. Select. Areas in Comm.*, Vol. 1, No. 1, pp. 46-55, January 2002.

[7] Andersin, M., and Rosberg, Z., "Time variant power control in cellular networks," in *Proc. 7th IEEE Intl. Symp. on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Vol. 1, pp. 193-197, October 1996.

[8] Tsoukatos, K. P., "Power control in a mobility environment," in *Proc. IEEE Vehicular Technology Conf. (VTC)*, pp. 740-744, 1997.

[9] Leung, Y.-W., "Power control in cellular networks subject to measurement error," *IEEE Trans. on Comm.*, Vol. 44, No. 7, pp. 772-5, July 1996.

[10] Ulukus, S., and Yates, R. D., "Stochastic power control for cellular radio systems," *IEEE Trans. on Comm.*, Vol. 46, No. 6, pp. 784-98, June 1998.

[11] Leung, K. K., "A Kalman Filter method for power control in broadband wireless networks," in *Proc. of INFOCOM'99* , pp. 948-56, 1999.

[12] Gunnarsson, F., *Power control in cellular radio systems: analysis, design and estimation*, Ph.D. Dissertation, Dept. of Electrical Engineering, Linkoping University, Sweden, 2000.

[13] Stuber, G., L., *Principles of mobile communications*, Norwell, MA: Kluwer Academic Publishers, 1996.

[14] Cox, D. C., and Reudink, D. O., "Dynamic channel assignment in high capacity mobile communication systems," *Bell Syst. Tech. J.*, pp. 1833-57, August 1971.

[15] Goodman, D. J., Grandhi, S. A., and Vijayan, R., "Distributed dynamic channel assignment schemes," in *Proc. IEEE Vehicular Technology Conf. (VTC)*, , pp. 532-35, 1993.

[16] Katzela, I., and Naghshineh, M., "Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey," *IEEE Personal Communications Magazine*, Vol. 3, pp. 10-31, June1996.

[17] Chuang, J., C.-I., and Sollenberger, N. R., "Performance of autonomous dynamic channel assignment and power control for TDMA/FDMA wireless acces," *IEEE J. Select. Areas in Comm.*, Vol. 12, No. 8, pp. 1314-23, October 1994.

[18] Lozano, A., and Cox, D. C., "Integrated dynamic channel assignment and power control in TDMA mobile wireless communication systems," *IEEE J. Select. Areas in Comm.*, Vol. 17, No. 11, pp. 2031-40, November 1999.

[19] Verdone, R., and Zanella, A., "On the optimization of fully distributed power control techniques in cellular radio systems," *IEEE Trans. on Vehicular Technology*, Vol. 49, No. 4, pp. 1440-48, July 2000.

[20] Foschini, G. J., and Miljanic, Z., "Distributed autonomous wireless channel assignment algorithm with power control," *IEEE Trans. on Vehicular Technology*, Vol. 44 , No. 3 , pp. 420-9, August 1995.

[21] Bambos, N. D., Chen, S. C., and Pottie, G. J., "Radio link admission algorithms for wireless networks with power control and active link quality protection," Technical Report UCLA-ENG-94-25, Elect. Eng. Dept., Univ. of California, Los Angeles, 1994.

[22] Bambos, N. D., Chen, S. C., and Mitra, D., "Channel probing for distributed access control in wireless communication networks," in *Proc. GLOBECOM'95*

[23] Wang, C. C-Y., *Power control strategies and variable bit allocation for FH-CDMA wireless systems*, Ph.D. Dissertation, Electrical Engineering Dept., UCLA, 1996.

[24] Hansen, C. J., *Probing techniques for multiuser channels with power control*, Ph.D. Dissertation, Electrical Engineering Dept., UCLA, 1997.

[25] Hansen, C. J., and Pottie, G. J., "A distributed access algorithm for cellular radio systems with channel partitioning," *IEEE Trans. on Vehicular Technology*, Vol. 48, No. 1, pp. 76-82, January 1999.

[26] Shoarinejad, K., Paganini, F., Pottie, G. J., and Speyer, J. L., "Global stability of feedback power control algorithms for cellular radio networks," in *Proceedings of the 40th IEEE Conf. on Decision and Control*, Vol. 1, pp. 610-615, 2001.

[27] Song, L., Mandayam, N. B., and Gajic, Z., "Analysis of an up/down power control algorithm for the CDMA reverse link under fading," *IEEE J. Select. Areas in Comm.*, Vol. 19, No. 2, pp. 277-286, February 2001.

[28] El-Osery, A. and Abdallah, C., "Distributed power control in CDMA cellular systems," *IEEE Antennas and Propagation Magazine*, Vol. 42, No. 4, pp. 152-9, August 2000.

[29] Jakes, W.C., *Microwave mobile communications*, New York, NY: John Wiley & Sons, 1974.

[30] Gudmundson, M., "Correlation model for shadow fading in mobile radio systems," *Electronics Letters*, Vol. 27, No. 23, pp. 2145-6, November 1991.

[31] Goldsmith, A. J., Greenstein, L. J., and Foschini, G. J., "Error statistics of real-time power measurements in cellular channels with multipath and shadowing," *IEEE Trans. on Vehicular Technology*, Vol. 43, No. 3, pp. 439-46, August 1994.

[32] Narasimhan, R., and Cox, D. C., "Speed estimation in wireless systems using wavelets," *IEEE Trans. on Comm.*, Vol. 47, No. 9, pp. 1357-64, September 1999.

[33] Sorenson, H. W., Ed., *Kalman filtering: theory and application*, New York, NY: IEEE Press, 1985.

[34] Pora, W., Chambers, J. A., and Constantinides, A. G., "Combination of Kalman filter and constant modulus algorithm with variable step-size for equalization of fast fading channels," *Electronics Letters*, Vol. 34, No. 18, pp. 1718-19, September 1998.

[35] Dziong, Z., Jia, M., and Mermelstein, P., "Adaptive traffic admission for integrated services in CDMA wireless-access networks," *IEEE J. Select. Areas in Comm.*, Vol. 14, No. 9, pp. 1737-47, December 1996.

[36] Ludwin, W., and Chlebus, E., "Performance comparison of simulators for cellular mobile networks," *Applied Mathematical Modeling*, Vol. 20, No. 8, pp. 585-7, Elsevier, August 1996.

[37] Yates, R. D., and Ching-Yao Huang, "Integrated power control and base station assignment," *IEEE Trans. on Vehicular Technology*, Vol. 44, No. 3, pp. 638-44, August 1995.

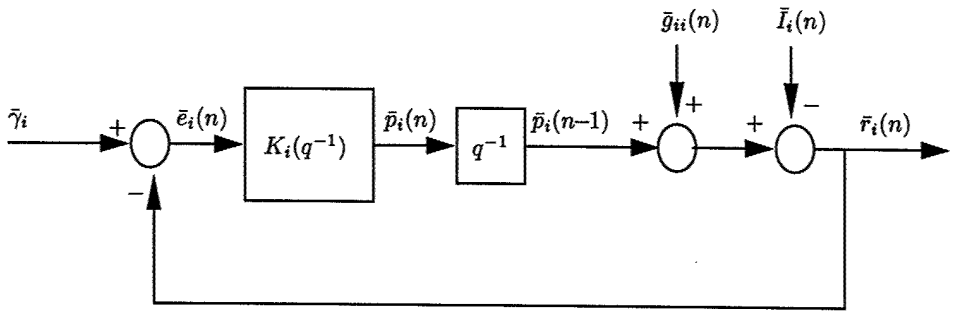


Figure 1. A local power control loop, associated with a single user

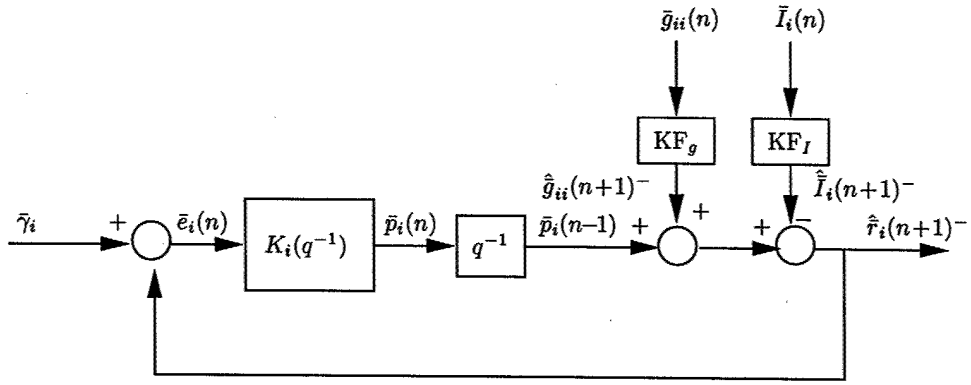


Figure 2. A local power control loop with Kalman filters

	Fixed-Step DCPA, 250 Hz	Predictive DCPA, 50 Hz
8.0 Erlang/Cell	0.17%	0.16 %
9.0 Erlang/Cell	0.73 %	0.66 %

Table 1

CALL DROPPING PERCENTAGE

	Fixed-Step DCPA, 250 Hz	Predictive DCPA, 50 Hz
8.0 Erlang/Cell	1.12%	0.86%
9.0 Erlang/Cell	3.27%	3.15%

Table 2

CALL BLOCKING PERCENTAGE

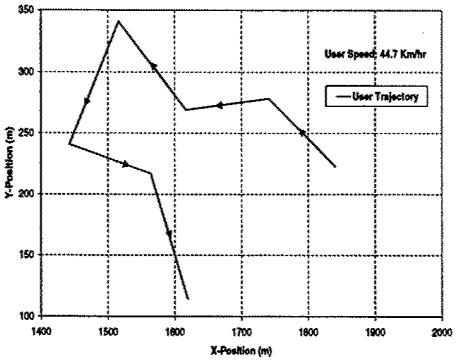


Figure 3. A Sample User Motion Trajectory

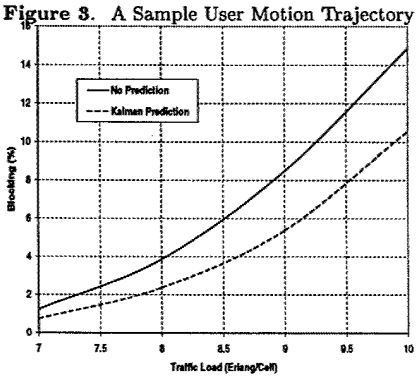


Figure 4. Call Blocking Response

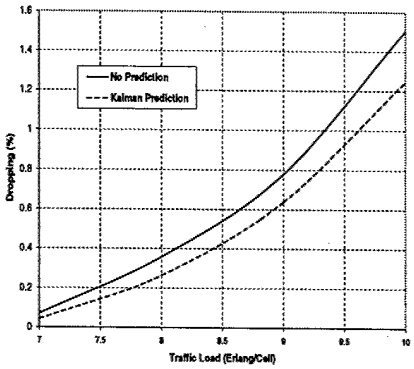


Figure 5. Call Dropping Response

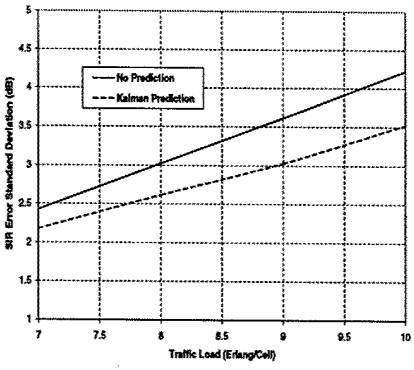


Figure 6. Standard Deviation for the Error in the Local Mean SIR

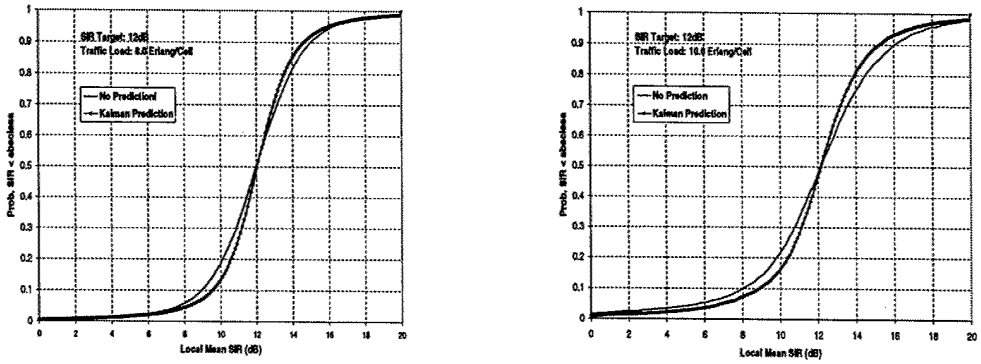


Figure 7. Cumulative Distributions for Local Mean SIR

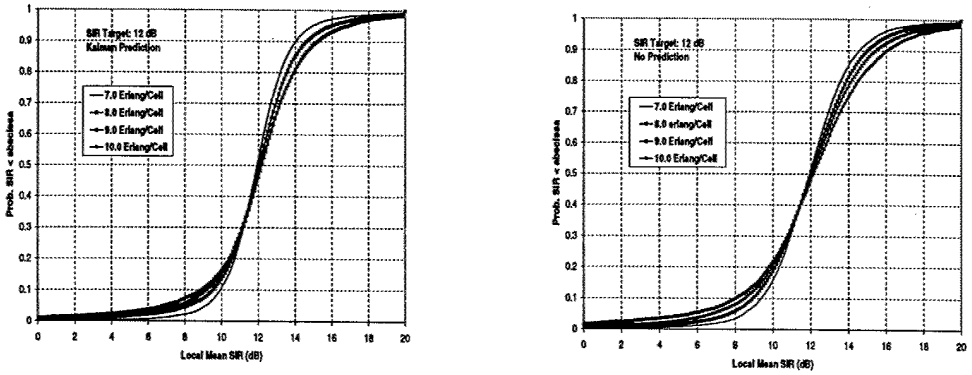


Figure 8. Traffic Load Effect on Local Mean SIR Cumulative Distribution

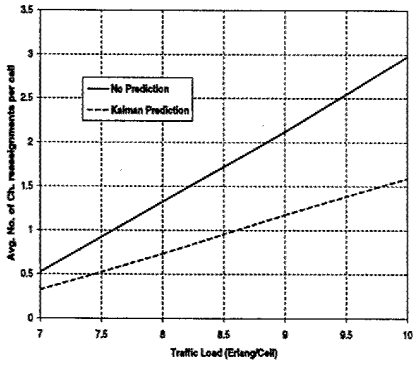


Figure 9. Average Number of Channel Reassignments per Call

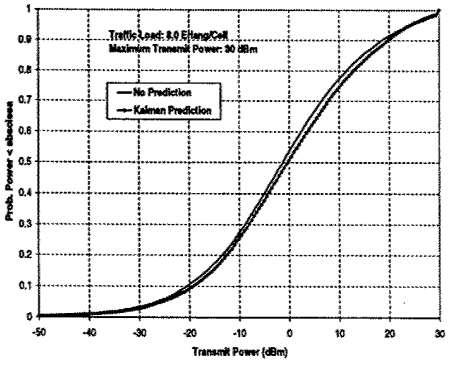


Figure 10. Transmit Power Cumulative Distribution at 8.0 Erlang/Cell

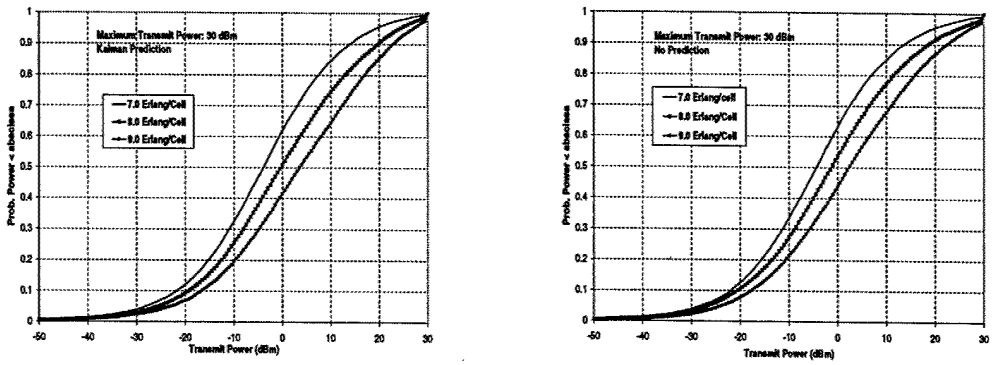


Figure 11. Traffic Load Effect on Transmit Power Cumulative Distribution

Appendix C

"Global Stability of Feedback Power Control Algorithms for Cellular Radio Networks"

Kambiz Shoarinejad, Jason L. Speyer, Fernando Paganini, and Gregory J. Pottie,

Proceedings of the IEEE CDC'01.

Global Stability of Feedback Power Control Algorithms for Cellular Radio Networks*

Kambiz Shoarinejad
Fernando Paganini
Gregory J. Pottie
UCLA Electrical Engineering
kambiz@ee.ucla.edu, paganini@ee.ucla.edu
pottie@ee.ucla.edu

Jason L. Speyer†
UCLA Mechanical and Aerospace Engineering
Box 951597, Los Angeles, CA 90095-1597
speyer@seas.ucla.edu

Abstract

Power control is considered as an efficient scheme to mitigate co-channel and multiple-access interference in cellular radio systems. Various approaches have been proposed in recent years to design power control algorithms. We focus on the feedback algorithms that are based on Signal to Interference plus Noise Ratios (SIR-based algorithms). We review SIR threshold approach and then discuss how power control design can be formulated as a decentralized regulation problem. We use a robust control framework to analyze global stability of a network on a single channel. We obtain a sufficient condition, which guarantees that the deviations of the power levels from their optimal values remain bounded, even when the channel gains change, as long as the network stays feasible.

1 Introduction

Optimal allocation of transmit power levels in wireless networks has attracted a lot of attention in recent years. The main idea is to control the transmit power level of a user or a base station in a wireless system in order to maintain an acceptable level of quality of service, while eliminating unnecessary interference to other users in the network. Different objectives and approaches have been perceived for power control and different algorithms have been naturally obtained.

The major objective in Direct Sequence Code Division Multiple Access systems is to mitigate the multiple access interference and therefore the near-far effect, whereas in Time/Frequency Division Multiple Access systems the objective is mostly to control the co-channel interference. Power control will also minimize the power consumption for the users and hence prolong their battery life.

We focus on power control algorithms that are based on Signal to Interference plus Noise Ratio (SIR). Note that, in general, higher SIR would yield better bit error performance and it is therefore common to abstract the system architecture and consider SIR as the measure for quality of service in order to formulate the power control objective.

One approach for SIR-based power control design is *SIR threshold* approach, presented in [1], where the objective is for the SIR of each user in the network to be above a desired threshold. It is shown how the optimal powers could be obtained through a simple distributed algorithm. The necessary and sufficient condition for the existence of the optimal powers is expressed as a *feasibility* condition. Various generalizations of this algorithm were later discussed in the literature. A uniform framework along with convergence analysis (under the condition of feasibility) for many of these algorithms were presented in [2].

In this paper, we focus on the decetralized regulator formulation for power control design. It has been noticed that the distributed algorithm presented in [1] is simply an integrator algorithm, in a closed loop, on the logarithmic scale. This has initiated a the decentralized regulator approach for power control design where concepts and design methodologies from control theory have been used for the analysis of current algorithms [3] and design of new algorithms [4][5]. This approach could be especifically helpful if models for fading, i.e., channel gain variations, are to be incorporated in the design. However, stability and

*This research was supported in part by the Air Force Office of Scientific Research under Grant Number F49620-00-1-0154
†Author to whom correspondence should be addressed.

convergence of these algorithms can not be verified through simple techniques such as the one presented in [2]. Therefore more complicated stability analysis should be performed to ensure *global stability* of the network under these power control algorithms. A robust control framework was presented in [4], where a sufficient condition for global stability was established using a linearized interference function. We use a similar framework to obtain another sufficient condition for global stability without any interference linearization. This condition will guarantee that, under a designed power control algorithm, the deviations of the power levels in the network from their corresponding optimal values will always remain bounded even when the channel gains change, as long as the variations in the channel gains do not force the network out of its feasibility region.

The organization of the paper is as follows. In the next section, we present the system model and review the SIR threshold approach. In Section 3, we review the decentralized regulator formulation for power control design, and in Section 4, we obtain a sufficient condition for global stability. Concluding remarks are provided in the final section.

2 System Model and SIR Threshold Approach

Consider a cellular system where M users are sharing a single channel. This channel could be a frequency band (FDMA), a time slot (TDMA) or even a spreading code (CDMA). Therefore, for every desired user-base station link, there are $M-1$ interfering links. The received SIR on the uplink channel for user i can now be written as:

$$r_i = \frac{g_{ii}p_i}{\sum_{j\neq i} g_{ij}p_j + \eta_i} \tag{1}$$

where p_i is the transmit power for user i, g_{ii} is the channel gain (or attenuation) from user i to its desired base station (in the linear scale), g_{ij} is the channel gain from user j to the desired base station of user i and η_i is the receiver noise intensity at the desired base station of user i. Note that even though we choose to focus on the uplink channel, a similar model and similar results can be obtained for the downlink channel. Define the normalized channel gain matrix Z as:

$$Z = [z_{ij}], \qquad z_{ij} = \frac{g_{ij}}{g_{ii}} \tag{2}$$

Note that Z is a non-negative stochastic matrix and, in general, is not symmetric.

In the SIR threshold approach, the objective is for the SIR of every user i to be above a desired threshold γ_i, that is: $r_i \geq \gamma_i$. It is easy to show that these constraints can be written in the matrix form as:

$$P \geq \Gamma(Z-I)P + U \tag{3}$$

where $\Gamma = diag(\gamma_1, \ldots, \gamma_M)$ and $U = [u_i] = [\frac{\gamma_i \eta_i}{g_{ii}}]$ and I is the identity matrix. The necessary and sufficient condition for the existence of a positive power vector P, which satisfies the above constraint, is called *feasibility*. In other words, a network of users is called feasible if every user can achieve its desired SIR. The corresponding power vector is then called a feasible power vector. It is clear that feasibility of a network depends on all channel gains and all desired SIRs. In SIR threshold approach, the feasibility condition is quantified and the minimum feasible power vector is obtained.

Theorem 2.1 (SIR Threshold) *Assuming $U > 0$, a network of users is feasible if and only if $\rho(F) < 1$, where:*

$$F \triangleq \Gamma(Z-I) \Rightarrow f_{ii} = 0, \qquad f_{ij} = \frac{\gamma_i g_{ij}}{g_{ii}}, \quad i \neq j \tag{4}$$

and ρ denotes the spectral radius of a matrix. Under the feasibility condition, the optimal power vector is obtained as:

$$P^* = (I-F)^{-1}U \tag{5}$$

Matrix F is a non-negative (component-wise) irreducible matrix and the above theorem can be proved using some results from the theory of non-negative matrices [6]. The power vector P^* is optimal in the sense that for any other feasible power vector P, we have $P > P^*$.

The above solution for P^* is a centralized solution in the sense that a central processor needs to gather all the information about all the channel gains and target SIRs, calculate the optimal power vector and send back the corresponding power command to every user. It was shown in [1] that a simple iterative algorithm, which could be implemented in a distributed manner, would converge to P^*. In fact, it is clear that under the condition of feasibility, the optimal power vector P^* is the unique fixed point of the following iterative algorithm:

$$P(n) = FP(n-1) + U \tag{6}$$

and component-wise, we can write:

$$\begin{aligned} p_i(n) &= \frac{\gamma_i}{g_{ii}}\left(\sum_{j\neq i} g_{ij}p_j(n-1) + \eta_i\right) \\ &= \frac{\gamma_i}{g_{ii}} I_i(n) = p_i(n-1)\frac{\gamma_i}{r_i(n)} \end{aligned} \tag{7}$$

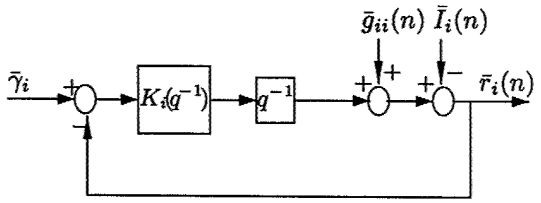


Figure 1: A local power control loop, associated with a single user

where $I_i(n)$ is the total interference plus noise power at the receiver of the assigned base station for user i. Therefore, at the beginning of the n-th power update period, the local mean channel gain g_{ii} and the local mean total interference plus noise power $I_i(n)$ are measured at the receiver and the new power level $p_i(n)$ is calculated and sent back to the user. Note that no extra delays are assumed for processing and propagation. Moreover, the convergence is proved assuming that all the channel gains and the desired SIRs stay constant for the duration of convergence of the algorithm. This may not always be a reasonable assumption, especially if fast fading is considered while low power update rates are assumed. In the next section, we discuss how power control can be posed as a decentralized regulator problem.

3 Decentralized Regulator Approach

Using a bar on the variables to indicate the values in dB, we can write the distributed algorithm in (2) in logarithmic scale as:

$$\bar{p}_i(n) = \bar{p}_i(n-1) + (\bar{\gamma}_i - \bar{r}_i(n)) \triangleq \bar{p}_i(n-1) + \bar{e}_i(n) \quad (8)$$

where $\bar{p}_i(n)$ is the power level in dBm for user i for the duration of the n-th power update period and $\bar{r}_i(n)$ is the SIR in dB for the same user at the beginning of the n-th power update period:

$$\bar{r}_i(n) = \bar{p}_i(n-1) + \bar{g}_{ii}(n) - \bar{I}_i(n) \quad (9)$$

Moreover, $\bar{I}_i(n)$ is the local mean interference plus noise power in dBm available at the beginning of the n-th power update period:

$$\bar{I}_i(n) = 10\log_{10}\left(\sum_{j\neq i} g_{ij} 10^{\frac{\bar{p}_j(n-1)}{10}} + \eta_i\right) \quad (10)$$

It is now easy to see that this algorithm is, in fact, a simple unity gain integrator algorithm in a closed local loop, as shown in Figure 1. The controller transfer function in this case is:

$$K_i(q^{-1}) = \frac{\bar{P}_i(q^{-1})}{\bar{E}_i(q^{-1})} = \frac{1}{1-q^{-1}} \quad (11)$$

where q is the shift operator. Therefore, the network can be seen as a set of interconnected local loops, each of which is associated with a single user. It should be realized that the couplings among the local loops is through the interference function (10), which, in general, is nonlinear. The decentralized regulator formulation of the power control problem can now be presented as: *Design a set of local controllers $K_i(q^{-1})$ such that the SIR for every user, $\bar{r}_i$, tracks a desired threshold $\bar{\gamma}_i$ with a certain performance while the global network remains stable.*

This approach has already initiated research on using control theory concepts for power control design [3]-[5]. Note that the local loops in Figure 1 are quite general and can be modified to accommodate different modeling assumptions. For example, extra delay blocks may be inserted in the feedback path to model processing and propagation delays. In fact, one step delay is typically assumed when high power update rates are considered [7]. It should also be mentioned that we have implicitly assumed a linear time invariant controller by writing $K_i(q^{-1})$. However, in general, the controller itself can be a nonlinear block, as is the case for *Fixed-Step* power control algorithms.

4 Global Stability

Unfortunately, stability and convergence of the power control algorithms, designed as decetralized regulation algorithms, can not be verified through simple techniques such as the one presented in [2]. A robust control framework was proposed in [4] to obtain a sufficient condition for global stability using a linearized interference function. We will use a similar approach, but with a different notion for stability, and we obtain a more general sufficient condition for global stability without any interference linearization.

We consider a system to be stable if bounded inputs generate bounded outputs. In robust control terminology [8][9], we use ℓ_∞ norm to quantify the size of the signals in the system and ℓ_∞-induced norms to quantify the amplification of the signals, i.e., the size of operators or transfer functions. We will obtain a sufficient global stability condition using a fundamental stability result called the *Small Gain Theorem*:

Theorem 4.1 (Small Gain Theorem) *Consider the feedback loop in Figure 2. Let $G_1 : \ell_\infty^n \to \ell_\infty^m$ and*

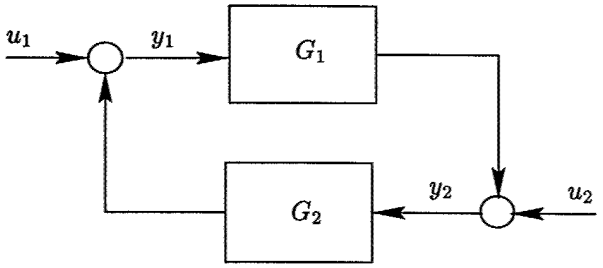


Figure 2: Closed-Loop Stability

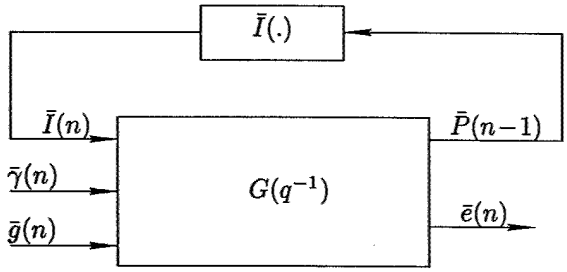


Figure 3: The Power-Controlled Global Network (on a single channel)

$G_2 : \ell_\infty^m \to \ell_\infty^n$ be two stable operators and assume that the closed loop system is well-posed (i.e., for any $u_1, u_2 \in \ell_\infty$, there exists a unique solution for $y_1, y_2 \in \ell_\infty$). Then the closed-loop system is stable if $||G_1||_{\ell_\infty - induced} ||G_2||_{\ell_\infty - induced} < 1$.

Note that the above theorem only states a sufficient condition, which may be conservative in some cases.

As we mentioned, a network of users can be seen as a nonlinearly coupled set of local loops. In fact the global network can be depicted as in Figure 3, where $G(q^{-1}) = \text{diag}\left(G_1(q^{-1}), \ldots, G_M(q^{-1})\right)$ is a block diagonal closed-loop transfer function matrix from interference $\bar{I}(n)$ to power $\bar{P}(n-1)$ and $\bar{I}(.)$ is a nonlinear operator, which produces interference plus noise power in dBm from the power levels. Note that $G_i(q^{-1})$ is also equal to the closed-loop transfer function from $\bar{\gamma}_i$ to $\bar{r}_i$. We have:

$$\bar{P}(n-1) = G(q^{-1})(\bar{I}(n) - \bar{g}(n) + \bar{\gamma}(n)) \qquad (12)$$

where:

$$\bar{g} \triangleq [\bar{g}_{11} \ldots \bar{g}_{MM}]^T \qquad (13)$$

$$\bar{\gamma} \triangleq [\bar{\gamma}_1 \ldots \bar{\gamma}_M]^T \qquad (14)$$

$$\bar{I}(n) \triangleq [\bar{I}_1(\bar{P}(n-1)) \ldots \bar{I}_M(\bar{P}(n-1))]^T \qquad (15)$$

Now lets assume that the network always stays feasible. Note that we are not assuming channel gains to be constant. But we only assume that the time variations of the channel gains do not push the network out of its feasibility region. Therefore, at any instant of time, there exists an instantaneous bounded optimal power vector $\bar{P}^*$, which is related to the corresponding optimal interference as:

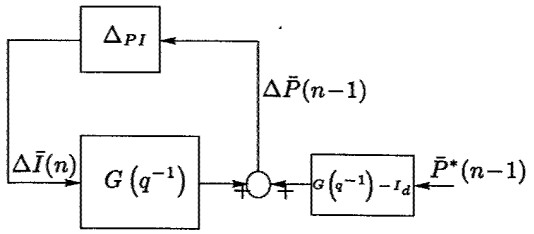


Figure 4: The Power-Controlled Global Network in a Variational Form

$$\bar{P}^*(n-1) = \bar{I}^*(n) - \bar{g}(n) + \bar{\gamma}(n) \qquad (16)$$

Since we are not considering user arrival or departures, $\bar{P}^*$ will be constant as long as the desired SIR thresholds and the channel gains remain constant. We now consider the deviations of the power and interference levels in the network, at every instant of time, relative to their optimal values, that is:

$$\Delta\bar{P} \triangleq \bar{P} - \bar{P}^* \qquad (17)$$

$$\Delta\bar{I} \triangleq \bar{I} - \bar{I}^* \qquad (18)$$

Using (12) and (16), we can now write:

$$\bar{P}(n-1) - G(q^{-1})\bar{P}^*(n-1) = G(q^{-1})\Delta\bar{I}(n) \qquad (19)$$

Hence:

$$\begin{aligned}\Delta\bar{P}(n-1) &= \bar{P}(n-1) - \bar{P}^*(n-1) \\ &= \bar{P}(n-1) - G(q^{-1})\bar{P}^*(n-1) + \left(G(q^{-1}) - I_d\right)\bar{P}^*(n-1) \\ &= G(q^{-1})\Delta\bar{I}(n) + \left(G(q^{-1}) - I_d\right)\bar{P}^*(n-1) \qquad (20)\end{aligned}$$

where I_d is the identity matrix. The network can then be shown as in Figure 4, where Δ_{PI} is the nonlinear operator transforming $\Delta\bar{P}$ to $\Delta\bar{I}$. We can show that Δ_{PI} is a contractive operator in the sense that $||\Delta_{PI}||_{\ell_\infty - induced} < 1$. To do so, we use the *Mean Value Theorem* [10].

Lemma 4.2

$$||\Delta_{PI}||_{\ell_\infty - induced} < 1 \qquad (21)$$

Proof: Using (10), it is straightforward to show:

$$\frac{\partial \bar{I}_i}{\partial \bar{p}_j} = \begin{cases} 0 & i = j \\ \frac{g_{ij}p_j}{\sum_{k \neq i} g_{ik}p_k + \eta_i} & i \neq j \end{cases} \qquad (22)$$

Remember again that the variables without bar indicate values in linear scale. From the Mean Value Theorem, we know that for every i and for every optimal power vector $\bar{P}^*$, there exists a power vector $\bar{\tilde{P}}$ lying on the line segment between $\bar{P}$ and $\bar{P}^*$ such that:

$$\Delta\bar{I}_i = \left.\frac{\partial \bar{I}_i}{\partial \bar{P}}\right|_{\bar{P}=\bar{\tilde{P}}} \Delta\bar{P} \tag{23}$$

Now using (22) and (23) and assuming $||\Delta\bar{P}||_\infty \leq 1$, which then implies $|\Delta\bar{p}_i(k)| \leq 1$ for all $i = 1, \ldots, M$ and $k = 0, 1, \ldots$, we can write:

$$|\Delta\bar{I}_i(k)| = \left| \sum_{j\neq i} \frac{g_{ij}(k)\tilde{p}_j(k-1)}{\sum_{l\neq i} g_{il}(k)\tilde{p}_l(k-1)+\eta_i} \Delta\bar{p}_j(k-1) \right| \tag{24}$$

$$\leq \sum_{j\neq i} \frac{g_{ij}(k)\tilde{p}_j(k-1)}{\sum_{l\neq i} g_{il}(k)\tilde{p}_l(k-1)+\eta_i} |\Delta\bar{p}_j(k-1)| \tag{25}$$

$$\leq \sum_{j\neq i} \frac{g_{ij}(k)\tilde{p}_j(k-1)}{\sum_{l\neq i} g_{il}(k)\tilde{p}_l(k-1)+\eta_i} < 1 \tag{26}$$

Therefore: $||\Delta\bar{I}||_\infty = \sup_k \max_i |\Delta\bar{I}_i(k)| < 1$, and hence: $||\Delta_{PI}||_{\ell_\infty - induced} = \sup_{||\Delta\bar{P}||_\infty \leq 1} ||\Delta\bar{I}||_\infty < 1$. Note that $||\Delta_{PI}||_{\ell_\infty - induced} = 1$ if no receiver noise is considered for any of the receivers.

It is clear that stability of every local loop is a necessary (but not sufficient) condition for global stability. We are now ready to state a sufficient condition for global stability of the network.

Theorem 4.3 (Global Stability) *Consider the network in Figure 4. Assume that the network is always feasible, i.e., there always exists a bounded power vector P^* satisfying (16). Then the network is globally stable if for every user i:*

$$||G_i(q^{-1})||_{\ell_\infty - induced} \leq 1 \tag{27}$$

Proof: Since $G_i(q^{-1})$ always incorporates a delay, it is easy to see that the operator $\Delta_{PI}G$ is always strictly causal and hence the closed loop system in Figure 4 is always well-posed. Moreover, the feasibility assumption guarantees the existence of a bounded P^*. Therefore, if $||G_i(q^{-1})||_{\ell_\infty - induced} \leq 1$ for every user i, we will have $||G(q^{-1})||_{\ell_\infty - induced} \leq 1$ and using Lemma 4.2, the global stability of the network will be established simply by invoking the Small Gain Theorem.

The above theorem states that if the feasibility condition is not violated and if (27) is satisfied, then the deviations of the power levels of all the users in the network from their corresponding optimal values will always remain bounded. Even though the condition (27) is only sufficient and might be conservative in some cases, it can still help us design new stable algorithms and analyze the stability of current algorithms under channel gain variations. We will show this by an example.

But first, we want to compare our result with the one presented in [4]. It was shown in [4] that if the channel gains stay constant and if the network is feasible (i.e., a constant optimal power vector exists) and if the interference function is linearized around this optimal power vector, then a sufficient condition for global stability of the linearized network (in the $\ell_2 - induced$ norm sense) is:

$$||G_i(q^{-1})||_{\ell_2 - induced} = \sup_\omega |G_i(e^{j\omega})| \leq 1 \tag{28}$$

This means that if the power vector of the network deviates a little bit from the optimal power vector, and as long as all the channel gains stay constant, the power levels will asymptotically move back to their optimal values. In contrast, in deriving the sufficient condition (27), no constant channel gain assumption was made and no linearization was involved. However, the stability in $\ell_\infty - induced$ norm does not imply asymptotic convergence of the small power level deviations to zero. Instead, it implies that the deviations always remain bounded even if the optimal power vector changes due to the variations in the channel gains. Also (27) is sometimes more conservative, since we always have:

$$||G_i(q^{-1})||_{\ell_2 - induced} \leq ||G_i(q^{-1})||_{\ell_\infty - induced} \tag{29}$$

Example: Consider the integral algorithm in (8) with gain β, i.e.,: $\bar{p}_i(k) = \bar{p}_i(k-1) + \beta(\bar{\gamma}_i - \bar{r}_i(k))$, or in linear scale:

$$p_i(k) = p_i(k-1)\left(\frac{\gamma_i}{r_i(k)}\right)^\beta. \tag{30}$$

We have:

$$G_i(q^{-1}) = \frac{q^{-1}K_i(q^{-1})}{1+q^{-1}K_i(q^{-1})} = \frac{\beta q^{-1}}{1-(1-\beta)q^{-1}} \tag{31}$$

We should first note that for the local loops to be stable we need to have $\beta \in (0, 2)$. It is now easy to show that for $0 \leq \beta \leq 1$, we have:

$$||G_i(q^{-1})||_{\ell_2 - induced} = ||G_i(q^{-1})||_{\ell_\infty - induced} = 1.0\,, \tag{32}$$

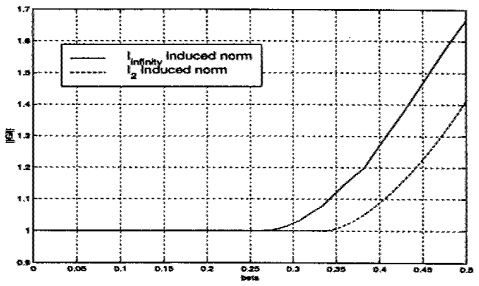


Figure 5: $\ell_\infty - induced$ and $\ell_2 - induced$ norms for G_i in the one step delayed case

and when β becomes larger than one, both induced norms start increasing. This proves that not only do the power levels, obtained from the distributed iterative algorithm in [1] (where $\beta = 1$ is assumed), converge to their optimal levels if the channel gains stay constant, but also, under the channel gain variations, the deviations of the power levels from their optimal values always remain bounded.

It is instructive to also consider the case where an additional delay is assumed for processing and propagation, i.e., one step delay is inserted in the feedback path in Figure 1. In this case:

$$G_i(q^{-1}) = \frac{q^{-2}K_i(q^{-1})}{1 + q^{-2}K_i(q^{-1})} = \frac{\beta q^{-2}}{1 - q^{-1} + \beta q^{-2}} \quad (33)$$

First note that $\beta = 1$ will result in closed-loop poles on the unit circle and therefore instability of the local loops. The $\ell_\infty - induced$ and $\ell_2 - induced$ norms of G_i are shown in Figure 5. It can be seen that in order to guarantee the bounded deviations of the power levels in the network (i.e., the global stability in the ℓ_∞ sense), we need to approximately have $\beta < .27$. Moreover, to ensure the global stability of the linearized system in the ℓ_2 sense, we need to have $\beta < 0.33$. These bounds on the gain are rather small and could therefore result in slow responses to the changes in SIR thresholds or the channel gains. However, remember that the sufficient conditions for global stability have been obtained under worst case scenarios and therefore might yield conservative requirements in some cases.

5 Conclusion

We reviewed SIR threshold approach for power control design in cellular wireless systems. Then we discussed the decentralized regulator formulation for power control problem. Using a robust control framework, we obtained a sufficient condition, which would guarantee that the deviations of the power levels from their corresponding optimal values always remain bounded. We then showed that if no extra delay is considered for processing and propagation, the widely proposed integrator algorithm does indeed yield a globally stable network as long as the variations of the channel gains do not force the network out of its feasibility region. As future work, we shall try to actually quantify some bounds on the power level deviations.

6 Acknowledgements

The authors would like to acknowledge helpful comments and suggestions from Professor Jeff Shamma, from UCLA MAE Department.

References

[1] Foschini, G. J. and Miljanic, Z.,"A simple distributed autonomous power control algorithm and its convergence", *IEEE Trans. on Vehicular Technology*, Vol. 42, No. 4, pp. 641-646, November 1993.

[2] Yates, R. D.,"A framework for uplink power control in cellular radio systems", *IEEE J. Select. Areas in Comm.*, Vol. 13, No. 7, pp. 1341-7, September 1995.

[3] Song, L., Mandayam, N. B., and Gajic, Z., "Analysis of an up/down power control algorithm for the CDMA reverse link under fading", Submitted to *IEEE J. Select. Areas in Comm.*, 2000.

[4] Gunnarsson, F., *Power control in cellular radio systems: analysis, design and estimation*, Ph.D. Dissertation, Dept. of Electrical Engineering, Linkoping University, Sweden, 2000.

[5] El-Osery, A. and Abdallah, C.,"Distributed power control in CDMA cellular systems", *IEEE Antennas and Propagation Magazine*, Vol. 42, No. 4, pp. 152-9, August 2000.

[6] Horn, R. A. and Johnson, C. R., *Matrix analysis*, Cambridge: Cambridge University Press, 1985.

[7] Viterbi, A., J., Viterbi, A. M., and Zehavi, E., " Performance of power-controlled wide-band terrestrial digital communication", *IEEE Trans. on Comm.*, Vol. 41, No. 4, pp. 559-569, April 1993.

[8] Dahleh, M. A., and Diaz-Bobillo, I. J., *Control of uncertain systems, a linear programming approach*, Englewood Cliffs, NJ: Prentice-Hall, 1995.

[9] Zhou, K., Doyle, J. C., and Glover, K., *Robust and optimal control*, Upper Saddle River, NJ: Prentice-Hall, 1995.

[10] Khalil, H. K., *Nonlinear systems*, Second Edition, Upper Saddle River, NJ: Prentice-Hall, 1996.

Appendix D

"The periodic optimality of LQ controllers satisfying strong stabilization,"

Jonathan D. Wolfe and Jason L. Speyer,

Proceedings of the IFAC workshop on periodic control, August, 2001 and to be publish in *Automatica.*

ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

automatica

PERGAMON

Automatica ■■ (■■■■) ■■–■■

www.elsevier.com/locate/automatica

The periodic optimality of LQ controllers satisfying strong stabilization☆

Jonathan D. Wolfe*, Jason L. Speyer

UCLA Mechanical and Aerospace Engineering Department, 48-121 Engineering IV P.O. Box 951597, Los Angeles, CA 90095-1597, USA

Received 3 July 2001; received in revised form 18 October 2002; accepted 26 April 2003

Abstract

An LQ strong stabilization problem is proposed. To determine when a controller with periodic gains is locally superior to a linear time invariant compensator for this problem, a Π test is presented. For systems with strictly proper transfer functions, it is proven that the frequency range where stable periodic controllers based on weak variations about the LTI case can give better performance than stable LTI compensators is finite. In the development, a means to evaluate the second partials of functions with respect to matrix-valued parameters is introduced. For those systems where periodic control is warranted, techniques for designing optimal periodic strongly stabilizing controllers are presented. Two examples detailing the application of the Π test are provided, as well as an optimal periodic controller design example.

Keywords: Optimal control; Chattering; Stability properties; LQG control; Periodic

1. Introduction

Often it is desired that an output feedback controller not only stabilize a plant, but be stable itself. The process of designing such a controller is referred to as the strong stabilization problem. It has recently been shown that all linear time-invariant (LTI) systems that are both detectable and controllable can be strongly stabilized by periodic controllers (Savkin & Petersen, 1998). The proposed controller design consists of a full state controller that during a period of length T operates without any measurements upon a propagated state estimate. At the end of the period, this state estimate is updated by a Luenberger estimator.

This method has some drawbacks, however. The period must be longer than a minimum length T_0 to ensure strong stabilization, and the gain of the controller between the periodic updates affects the size of T_0. Because a large period implies poor performance in the presence of disturbances, T_0 must be kept reasonably small, but reducing T_0 requires high controller gains. Also, it is worrisome from a robustness standpoint that the controller runs open loop over each period. We will demonstrate that the disturbance rejection capability of a stable continuous feedback controller is considerably better.

The primary contribution of this paper is a cost function formulation that induces strong stability. Because this cost is non-convex, it provides an opportunity for periodic strongly stabilizing controllers to produce a lower cost than strongly stabilizing LTI controllers. Before designing a strongly stabilizing controller, however, it is wise to investigate the following related question: If we restrict ourselves to considering only observer-structure controllers, and require the controller to be stable, when can a control with periodic gains potentially reduce the cost function compared to one with fixed gains? To answer this question, we construct a Π test (Bittanti, Fronza, & Guardabassi, 1973; Bernstein & Gilbert, 1980) that indicates when small periodic variations from the best time invariant controller may improve the cost function. Of interest in its own right is the procedure we develop for converting problems where the optimization parameters are gain matrices into a form amenable to application of the Π test. Since a considerable number of fixed structure problems (including the static output feedback problem and

☆ A portion of this paper was presented in August 2001 at the IFAC Workshop on Periodic Control Systems in Cernobbio-Como, Italy. This paper was recommended for publication in revised form by Associate Editor ■■■ under the direction of Editor ■■■.

* Corresponding author.

E-mail addresses: wolfe@talus.seas.ucla.edu, jwolfe@ucla.edu (J.D. Wolfe), speyer@seas.ucla.edu (J.L. Speyer).

0005-1098/03/$ - see front matter © 2003 Published by Elsevier Science Ltd.
doi:10.1016/S0005-1098(03)00178-X

several decentralized control problems) involve optimizing over gain matrices, the method derived here appears to have many extensions.

We then develop techniques for designing periodic controllers that minimize our cost function and thus satisfy strong stability. It should be emphasized that a periodic stable controller may be determined even when a static stable controller does not exist. Furthermore, although the Π test will indicate when an LTI stable compensator is not a local minimum (and that a stable periodic design can outperform it), failure of the Π test does not imply that the stable LTI design is globally optimal, so construction of a stable periodic controller may still be worthwhile.

This paper is organized as follows: Section 2 formulates a new cost function that penalizes unstable controllers. In Section 3, we state some results on the second derivatives of traces of matrix functions that are interesting in themselves for numerical optimization of fixed structure controllers. The Π test for strongly stabilizing controllers is presented in Section 5, and it is shown that when a plant transfer function is strictly proper, gains based on weak variations from the static gains can reduce the cost function below the cost with an LTI controller only over a finite frequency range. Section 6 describes the process of optimal periodic controller design. The Π test is applied to example systems and an optimal periodic controller is demonstrated in Section 7. Section 8 concludes the paper.

2. The optimal control problem with a strong stabilization constraint

Consider the Gauss–Markov time-invariant system described by

$$\mathrm{d}x = (Ax + Bu)\,\mathrm{d}t + [\Gamma_1\ 0]\,\mathrm{d}\tilde{w}, \tag{1}$$

$$\mathrm{d}y = (Cx + Du)\,\mathrm{d}t + [0\ \Gamma_2]\,\mathrm{d}\tilde{w}, \tag{2}$$

where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^p$, $u \in \mathbb{R}^m$ and $\tilde{w} \in \mathbb{R}^q$ is a Brownian motion process whose independent increment processes $\mathrm{d}\tilde{w}$ have the statistics

$$E[\mathrm{d}\tilde{w}\,\mathrm{d}\tilde{w}^{\mathrm{T}}] = I\,\mathrm{d}t, \qquad E[\mathrm{d}\tilde{w}] = 0, \tag{3}$$

where $E[\cdot]$ indicates the expectation operator and I indicates the identity matrix. Without loss of generality, Γ_2 is assumed to have full row rank. Our cost function is the expectation of the quadratic cost function suggested in Bittanti et al. (1973):

$$J[u,\tau] = \lim_{k\to\infty} \frac{1}{k\tau} E\left[\int_0^{k\tau} (x^{\mathrm{T}}Qx + u^{\mathrm{T}}Ru)\,\mathrm{d}t\right], \tag{4}$$

where τ is the period of a cycle, k is the number of cycles, Q is a symmetric nonnegative definite matrix and R is a symmetric positive definite matrix. The answer to this optimization problem is the well-known linear quadratic Gaussian controller

$$\mathrm{d}\hat{x} = A\hat{x}\,\mathrm{d}t + Bu\,\mathrm{d}t + L(\mathrm{d}y - C\hat{x} - Du)\,\mathrm{d}t, \tag{5}$$

$$u = -K\hat{x}, \tag{6}$$

where K is the linear-quadratic regulator gain and L is the gain of the Kalman–Bucy filter.

Observe that if we define $e \triangleq x - \hat{x}$, the closed loop dynamics and cost can be rewritten as

$$d\begin{bmatrix} x \\ e \end{bmatrix} = \begin{bmatrix} A - BK & BK \\ 0 & A - LC \end{bmatrix}\begin{bmatrix} x \\ e \end{bmatrix}\mathrm{d}t + \begin{bmatrix} \Gamma_1 & 0 \\ \Gamma_1 & -L\Gamma_2 \end{bmatrix}\mathrm{d}\tilde{w}, \tag{7}$$

$$z_{LQG} = \begin{bmatrix} Q^{1/2} & 0 \\ -R^{\frac{1}{2}}K & R^{1/2}K \end{bmatrix}\begin{bmatrix} x \\ e \end{bmatrix}, \tag{8}$$

$$J[K,L,\tau] = \lim_{k\to\infty} \frac{1}{k\tau} E\left[\int_0^{k\tau} (z_{LQG}^{\mathrm{T}} z_{LQG})\,\mathrm{d}t\right]. \tag{9}$$

Note that the dynamics of the controller are described by

$$A_{\mathrm{c}} \triangleq A - BK - LC + LDK, \tag{10}$$

and that this matrix need not be Hurwitz.

Suppose we were to add a cost term that would penalize an unstable controller. If we constrained the controller to have the same observer structure as before, the dynamics and cost would look like

$$\mathrm{d}x_{\mathrm{cl}} = A_{\mathrm{cl}}x_{\mathrm{cl}}\,\mathrm{d}t + B_{\mathrm{cl}}\,\mathrm{d}w, \tag{11}$$

$$z = C_{\mathrm{cl}}x_{\mathrm{cl}}, \tag{12}$$

$$J[K,L,\tau] = \lim_{k\to\infty} \frac{1}{k\tau} E\left[\int_0^{k\tau} (z^{\mathrm{T}}z)\,\mathrm{d}t\right] \tag{13}$$

with

$$E[\mathrm{d}w\,\mathrm{d}w^{\mathrm{T}}] = I\,\mathrm{d}t, \quad E[\mathrm{d}w] = 0, \quad x_{\mathrm{cl}}^{\mathrm{T}} = [x^{\mathrm{T}}e^{\mathrm{T}}x_{\mathrm{f}}^{\mathrm{T}}],$$

$$A_{\mathrm{cl}} = \begin{bmatrix} A - BK & BK & 0 \\ 0 & A - LC & 0 \\ 0 & 0 & A_{\mathrm{c}} \end{bmatrix},$$

$$B_{\mathrm{cl}} = \begin{bmatrix} \Gamma_1 & 0 & 0 \\ \Gamma_1 & -L\Gamma_2 & 0 \\ 0 & 0 & \Gamma_{\mathrm{f}} \end{bmatrix},$$

$$C_{\mathrm{cl}} = \begin{bmatrix} Q^{1/2} & 0 & 0 \\ -R^{1/2}K & R^{1/2}K & 0 \\ 0 & 0 & Q_{\mathrm{f}}^{1/2} \end{bmatrix}, \tag{14}$$

where the new state x_f, with the dynamics of the controller, is forced by noise and included in the cost function via the weight Q_f. To insure that all controller states are penalized in the cost, it is required that Q_f be positive definite and that Γ_f have full row rank.

The cost expression above can be written in terms of the covariance of x_{cl}, $P \triangleq [x_{cl}x_{cl}^T]$:

$$J = \lim_{k\to\infty} \mathbf{tr}\, \frac{1}{k\tau} \int_0^{k\tau} \{PC_{cl}(t)^T C_{cl}(t)\}\, dt, \tag{15}$$

subject to

$$\dot{P} = A_{cl}(t)P + PA_{cl}(t)^T + B_{cl}(t)B_{cl}(t)^T, \tag{16}$$

$$P > 0, \tag{17}$$

where **tr** denotes the trace operation. Let us partition P into $n \times n$ pieces as follows:

$$P = \begin{bmatrix} P_1 & P_{12} & P_{13} \\ P_{12}^T & P_2 & P_{23} \\ P_{13}^T & P_{23}^T & P_3 \end{bmatrix}. \tag{18}$$

An equivalent expression of the cost is then

$$J = \lim_{k\to\infty} \mathbf{tr}\, \frac{1}{k\tau} \int_0^{k\tau} \{P_1 Q + (P_1 - P_{12} - P_{12}^T + P_2)K^T R K + P_3 Q_f\}\, dt. \tag{19}$$

We can write a Hamiltonian for this optimization problem in the usual way

$$\mathcal{H} = \mathbf{tr}\{P_1 Q + (P_1 - P_{12} - P_{12}^T + P_2)K^T R K + P_3 Q_f + \Lambda(A_{cl}P + PA_{cl}^T + B_{cl}B_{cl}^T)\}, \tag{20}$$

where $\mathcal{H}$ is almost identical to the Hamiltonian used in Denham and Speyer (1964) and is similar to the one (for a case with no process noise) used in Athans (1968). Following the standard derivation, the necessary conditions for minimizing J are:

Theorem 1 (Pontryagin's necessary conditions). *The following are necessary for minimizing J:*

(1) $\mathcal{H}$ *is minimized with respect to K and L,*
(2) $\mathcal{H}_P = -d\Lambda/dt$, $\Lambda(k\tau) = 0$, *for* $k = 0, 1, 2, \ldots$,
(3) $\mathcal{H}_\Lambda = dP/dt$.

If $\mathcal{H}$ has a minimum and is continuously differentiable in K and L, a necessary condition for minimizing $\mathcal{H}$ is $\mathcal{H}_K = \mathcal{H}_L = 0$. If we partition Λ in the same manner as P was partitioned in (18) and assume that there is a steady-state stationary solution to the optimization problem, then the following equations are satisfied at the stationary point:

$$\mathcal{H}_K = 0 = 2RK(P_1 - P_{12} - P_{12}^T + P_2) + 2D^T L^T(\Lambda_{13}^T P_{13} + \Lambda_{23}^T P_{23} + \Lambda_3 P_3) + 2B^T(-\Lambda_1 P_1 + \Lambda_1 P_{12} - \Lambda_{12}P_{12}^T + \Lambda_{12}P_2 - \Lambda_{13}^T P_{13} - \Lambda_{13}P_{13}^T + \Lambda_{13}P_{23}^T - \Lambda_{23}^T P_{23} - \Lambda_3 P_3), \tag{21}$$

$$\mathcal{H}_L = 0 = 2\Lambda_2 L\Gamma_2\Gamma_2^T + 2(\Lambda_{13}^T P_{13} + \Lambda_{23}^T P_{23} + \Lambda_3 P_3)K^T D^T - 2(\Lambda_{12}^T P_{12} + \Lambda_{13}^T P_{13} + \Lambda_2 P_2 + \Lambda_{23}^T P_{23} + \Lambda_{23}P_{23}^T + \Lambda_3 P_3)C^T, \tag{22}$$

$$-d\Lambda/dt = 0 = A_{cl}P + PA_{cl}^T + B_{cl}B_{cl}^T, \tag{23}$$

$$dP/dt = 0 = A_{cl}^T\Lambda + \Lambda A_{cl} + C_{cl}^T C_{cl}. \tag{24}$$

2.1. When is there a steady-state stationary solution to the optimization problem

The conditions for determining when an LTI system may be stabilized by a stable controller were found in Youla, Bongiorno, and Lu (1974). The following extension of these conditions to the MIMO case can be found in Vidyasagar (1985):

Theorem 2 (Parity interlacing property). *Let* $\mathbb{C}_{+e}$ *denote the extended right half of the complex plane* ($\{s \in \mathbb{C}: \mathbf{Re}(s) \geqslant 0\}$, *together with positive infinity*). *A plant P is strongly stabilizable if and only if the number of poles of P (counted according to their McMillan degrees) between any pair of real* $\mathbb{C}_{+e}$*-blocking zeros of P is even.*

Note that the stable compensator that stabilizes the system in the above theorem is a proper matrix transfer function of arbitrary order—i.e. a strictly proper stabilizing compensator with the same order as the plant may not exist. However, there are constructive sufficient conditions for stable, strictly proper full-order compensation (Wang & Bernstein, 1994). If such a compensator can be found, it can be used as a starting point for an iterative scheme to find a stationary point of our optimization problem (Geromel & Bernussou, 1979; Toivonen & Mäkilä, 1985, 1987).

3. Some results on second-order derivatives of traces

The sequel will require some results on second-order derivatives of traces of matrix functions. The proofs are substantially the same for each case, so the proof for one representative case has been included in Appendix A. Each of the other assertions can be proven using similar arguments.

Proposition 3. *Let X, Y, A, B be complex matrices of appropriate dimension. Denote the (i, j)th component*

of a matrix by $(\,)_{ij}$. *Then*

- $\dfrac{\partial^2}{\partial y_{kl}\partial x_{ij}}\mathbf{tr}(XAYB) = b_{li}a_{jk}$.
- $\dfrac{\partial^2}{\partial y_{kl}\partial x_{ij}}\mathbf{tr}(XAY^{\mathrm{T}}B) = b_{ki}a_{jl}$.
- $\dfrac{\partial^2}{\partial y_{kl}\partial x_{ij}}\mathbf{tr}(XAY^{\mathrm{T}}BYC) = c_{li}[B^{\mathrm{T}}YA^{\mathrm{T}}]_{kj} + [BYC]_{ki}a_{jl}$.
- $\dfrac{\partial^2}{\partial x_{kl}\partial x_{ij}}\mathbf{tr}(XAX^{\mathrm{T}}B) = b_{ki}a_{jl} + b_{ik}a_{lj}$.
- $\dfrac{\partial^2}{\partial y_{kl}\partial x_{ij}}\mathbf{tr}(X^{\mathrm{T}}AYB) = a_{ik}b_{lj}$.
- $\dfrac{\partial^2}{\partial y_{kl}\partial x_{ij}}\mathbf{tr}(X^{\mathrm{T}}AY^{\mathrm{T}}B) = a_{il}b_{kj}$.
- $\dfrac{\partial^2}{\partial y_{kl}\partial x_{ij}}\mathbf{tr}(X^{\mathrm{T}}AY^{\mathrm{T}}BYC) = a_{il}[BYC]_{kj} + [B^{\mathrm{T}}YA^{\mathrm{T}}]_{ki}c_{lj}$.

4. Products that convert linear matrix equations into linear vector equations

4.1. Review of the Kronecker product

The following well-known results can be found in elementary linear algebra texts (e.g. Lancaster, 1969, Chapter 8):

Definition 4 (Kronecker operator). Let $\mathscr{F}$ denote a field. If $A \in \mathscr{F}_{m\times n}$ and $B \in \mathscr{F}_{o\times p}$, then the Kronecker operation on A and B, written $A \otimes B$, is an $mo \times np$ matrix whose elements are defined by the relation $[A\otimes B]_{kl} = a_{ri}b_{sj}$, where $k = (r-1)o + s$, $l = (i-1)p + j$.

Proposition 5 (Kronecker product). *If* $A \in \mathscr{F}_{m\times n}$ *and* $B \in \mathscr{F}_{o\times p}$, *then the Kronecker operation* $A \otimes B$ *is a well-defined product.*

Proposition 6 (Kronecker product and linear matrix equations). *Consider the following matrix linear equation for the unknown matrix* $X \in \mathscr{F}_{n\times n}$: $AXB = C$ *where* $A, B, C \in \mathscr{F}_{n\times n}$. *We can consider this equation as an abbreviation for* n^2 *scalar equations for the* n^2 *elements of* X. *Let us define the "vectorized" versions of* X *and* C *in* $\mathscr{F}_{n^2}$ *by*

$$\mathbf{x}^{\mathrm{T}} = [X_{1*}^{\mathrm{T}}\ X_{2*}^{\mathrm{T}}\ \cdots\ X_{n*}^{\mathrm{T}}]^{\mathrm{T}}, \qquad \mathbf{c}^{\mathrm{T}} = [C_{1*}^{\mathrm{T}}\ C_{2*}^{\mathrm{T}}\ \cdots\ C_{n*}^{\mathrm{T}}]^{\mathrm{T}},$$

where X_{i*}, C_{j*} *denote the* i*th row of* X *and the* j*th row of* C, *respectively. Then the equation* $AXB = C$ *is equivalent to* $G\mathbf{x} = \mathbf{c}$ *for some* $G \in \mathscr{F}_{n^2\times n^2}$. *One can easily verify that* $G = A \otimes B^{\mathrm{T}}$.

Proposition 7 (Kronecker product of positive definite matrices). *If* A *and* B *are two positive definite matrices, then* $A \otimes B$ *is also positive definite.*

4.2. Another product

Recall that the matrix equation $AXB = C$ can be transformed to the form $(A\otimes B^{\mathrm{T}})\mathbf{x} = \mathbf{c}$, where $\mathbf{x}$ and $\mathbf{c}$ "vectorize" X and C by rows. Now, suppose we wished to express the matrix equation $AX^{\mathrm{T}}B = C$ as $G\mathbf{x} = \mathbf{c}$, where $\mathbf{x}$ and $\mathbf{c}$ are the same as before. Motivated by this problem, we will define a new operator.

Definition 8 (KT-operator). Let $\mathscr{F}$ denote a field. If $A \in \mathscr{F}_{m\times n}$ and $B \in \mathscr{F}_{o\times p}$ then the KT-operation on A and B, written $A \overset{\mathrm{T}}{\otimes} B$, is defined element-wise by

$$[A \overset{\mathrm{T}}{\otimes} B]_{kl} = a_{rj}b_{si},$$

where $k = (r-1)o + s$ and $l = (i-1)n + j$.

Proposition 9 (KT-product). *If* $A \in \mathscr{F}_{m\times n}$ *and* $B \in \mathscr{F}_{o\times p}$, *then the KT-operation* $A \overset{\mathrm{T}}{\otimes} B$ *is a well defined product.*

Proposition 10. *Let* $C \in \mathscr{F}_{m\times n}$, $A \in \mathscr{F}_{m\times o}$, $X \in \mathscr{F}_{p\times o}$, $B \in \mathscr{F}_{p\times n}$. *Let* $AX^{\mathrm{T}}B = C$ *be a linear matrix equation in* X. *"Vectorize"* X *and* C *as follows*:

$$\mathbf{x}^{\mathrm{T}} = [X_{1*}^{\mathrm{T}}\ X_{2*}^{\mathrm{T}}\ \cdots\ X_{n*}^{\mathrm{T}}]^{\mathrm{T}}, \qquad \mathbf{c}^{\mathrm{T}} = [C_{1*}^{\mathrm{T}}\ C_{2*}^{\mathrm{T}}\ \cdots\ C_{n*}^{\mathrm{T}}]^{\mathrm{T}}.$$

Then $AX^{\mathrm{T}}B = C$ *is equivalent to the equation* $(A \overset{\mathrm{T}}{\otimes} B^{\mathrm{T}})\mathbf{x} = \mathbf{c}$.

The proofs of the above propositions are trivial modifications of the corresponding proofs for the Kronecker product case.

Remark 11 (Relationship between Kronecker product and KT product). If X is a matrix and the column vector $\mathbf{x}$ is X vectorized by columns, then there exists a permutation matrix S, whose elements are all either 0 or 1, such that $S\mathbf{x}$ is X^{T} vectorized by columns. Then $AX^{\mathrm{T}}B = C$ is equivalent to both $(A \otimes B^{\mathrm{T}})S\mathbf{x} = \mathbf{c}$ and $(A \overset{\mathrm{T}}{\otimes} B^{\mathrm{T}})\mathbf{x} = \mathbf{c}$.

Using the KT product is preferable to using the Kronecker product and a permutation for two reasons: the notation is more compact, and the operation count is lower (the operation count for computing a KT product is the same as that required for a Kronecker product, while multiplying permutation matrices is costly due to the large size of the matrix).

5. Constructing a Π test

Before the Π test can be constructed, we must first obtain expressions for the partial derivatives of $\mathscr{H}$ and for the linearized equations of motion.

5.1. Partial derivatives of $\mathscr{H}$

Recall the Hamiltonian for our optimization problem is

$$\mathscr{H} = \mathbf{tr}\{P_1 Q + (P_1 - P_{12} - P_{12}^{\mathrm{T}} + P_2)K^{\mathrm{T}}RK + P_3 Q_f + \Lambda(A_{\mathrm{cl}}P + PA_{\mathrm{cl}}^{\mathrm{T}} + B_{\mathrm{cl}}B_{\mathrm{cl}}^{\mathrm{T}})\}. \quad (25)$$

Note that since P appears linearly in $\mathscr{H}$, $\partial^2\mathscr{H}/(\partial p_{kl}\partial p_{ij}) = 0$ $\forall i,j,k,l$.

The second partials of $\mathscr{H}$ can be found using the results on second partials of traces with respect to matrices developed in Section 3. These results can be expressed in a more compact notation if we "vectorize" the parameter matrices and write our results in terms of Kronecker products and KT-products. As an illustrative example, we will derive the second partial of $\mathscr{H}$ with respect to K.

Using the formulas in Section 3, one can find that

$$\frac{\partial^2}{\partial k_{gh}\partial k_{ef}}\mathscr{H}(P,K,L,\Lambda) = R_{ge}[P_1 - P_{12} - P_{12}^{\mathrm{T}} + P_2]_{fh} + R_{eg}[P_1 - P_{12} - P_{12}^{\mathrm{T}} + P_2]_{hf}. \quad (26)$$

Let δK be a small variation in K. Vectorize δK by rows, i.e. $\delta\mathbf{k}_{(q-1)n+r} = \delta K_{qr}$. Define $\mathscr{H}(X,K,L,\Lambda)_{\mathbf{kk}}$ as follows:

$$\delta\mathbf{k}^{\mathrm{T}}\mathscr{H}(P,K,L,\Lambda)_{\mathbf{kk}}\delta\mathbf{k} = \sum_{g=1}^{m}\sum_{h=1}^{n}\sum_{e=1}^{m}\sum_{f=1}^{n}\delta K_{gh}\frac{\partial^2\mathscr{H}(P,K,L,\Lambda)}{\partial k_{gh}\partial k_{ef}}\delta K_{ef}. \quad (27)$$

Then, using what we know about Kronecker products and Eqs. (27) and (26) we have

$$\mathscr{H}(P,K,L,\Lambda)_{\mathbf{kk}} = 2R \otimes [P_1 - P_{12} - P_{12}^{\mathrm{T}} + P_2]. \quad (28)$$

In an entirely analogous way, we can define $\mathbf{l}, \mathbf{p}_1, \mathbf{p}_{12}, \mathbf{p}_{13}, \mathbf{p}_2, \mathbf{p}_{23}, \mathbf{p}_3$ with respect to L and P. Then $\mathscr{H}_{\mathbf{kl}}$, $\mathscr{H}_{\mathbf{kp}_1}$, $\mathscr{H}_{\mathbf{kp}_{12}}$, $\mathscr{H}_{\mathbf{kp}_{13}}$, $\mathscr{H}_{\mathbf{kp}_2}$, $\mathscr{H}_{\mathbf{kp}_{23}}$, $\mathscr{H}_{\mathbf{kp}_3}$, $\mathscr{H}_{\mathbf{ll}}$, $\mathscr{H}_{\mathbf{lp}_1}$, $\mathscr{H}_{\mathbf{lp}_{12}}$, $\mathscr{H}_{\mathbf{lp}_{13}}$, $\mathscr{H}_{\mathbf{lp}_2}$, $\mathscr{H}_{\mathbf{lp}_{23}}$, $\mathscr{H}_{\mathbf{lp}_3}$ can be determined in terms of Kronecker and KT-products of the system matrices. These expressions are given in Appendix B.

5.2. Linearization of the equation of motion

The covariance P satisfies the differential equation

$$\dot{P}(t) = A_{\mathrm{cl}}(t)P(t) + P(t)A_{\mathrm{cl}}(t)^{\mathrm{T}} + B_{\mathrm{cl}}(t)B_{\mathrm{cl}}(t)^{\mathrm{T}}. \quad (29)$$

To linearize this bilinear form, suppose that P^0, K^0, L^0 are nominal solutions that satisfy (29). Then take small variations so that $P = P^0 + \delta P, K = K^0 + \delta K, L = L^0 + \delta L$. We can eliminate the higher-order terms and express the result in terms of "vectorized" quantities. This is easily accomplished using the rules for "vectorizing" matrix equations given in the sections discussing the Kronecker and KT products. For instance,

$$\begin{aligned}\delta\dot{\mathbf{p}}_1 &= [(A - BK^0)\otimes I + I\otimes(A - BK^0)]\delta\mathbf{p}_1 \\ &\quad + [(BK^0)\overset{\mathrm{T}}{\otimes} I + I\otimes(BK^0)]\delta\mathbf{p}_{12} \\ &\quad + [-B\otimes P_1^0 + B\otimes P_{12}^0 - P_1^0\overset{\mathrm{T}}{\otimes} B + P_{12}^0\overset{\mathrm{T}}{\otimes} B]\delta\mathbf{k}.\end{aligned} \quad (30)$$

The state space equations for this and the other $\delta\mathbf{p}_i$'s (which are given in Appendix C) can be put together into a large linear system

$$\delta\dot{\mathbf{p}} = \bar{F}\delta\mathbf{p} + \bar{G}[\delta\mathbf{k}\ \delta\mathbf{l}]. \quad (31)$$

A transfer function from the parameter variations $\delta\mathbf{k}$ and $\delta\mathbf{l}$ to the states $\delta\mathbf{p}$ can then be computed in the standard way:

$$\delta\mathbf{p}(s) = (sI - \bar{F})^{-1}\bar{G}[\delta\mathbf{k}(s)\ \delta\mathbf{l}(s)]. \quad (32)$$

5.3. The Π test

We will now create a Π test for the fixed structure strong stabilization problem, following the same general strategy used in the state feedback case (Bittanti et al., 1973; Bernstein & Gilbert, 1980). Consider nonlinear system (16) and associated cost (15). Let (31) be the linearization of the dynamics described in (16). Suppose also that we have found a set of static control and observer gains that meet the first-order necessary conditions for optimality.

Definition 12. An optimal periodic control problem is said to be *proper* if there exists a period $\hat{\tau}$ and an admissible control gains $\hat{K}(t), \hat{L}(t)$ such that

$$J[\hat{K}(t), \hat{L}(t), \hat{\tau}] < \bar{J}^0, \quad (33)$$

where $\bar{J}^0$ is the cost corresponding to the optimal steady-state solution of the problem, using the static gains $\bar{K}^0, \bar{L}^0$. Hence, a strong variation in the controller gains from the steady-state solution has a lower cost.

Note that the term "proper" has historically been used both to describe the optimality of periodic optimal control problems and to describe transfer functions that have more poles than zeros. To avoid confusion, we shall always explicitly state whether it is a periodic optimal control problem or a transfer function that is proper.

Definition 13. An optimal periodic control problem is said to be *locally proper* if there exists a period $\hat{\tau}$ and admissible weak variations $\hat{\delta}K(t), \hat{\delta}L(t)$ in the controller gains such that

$$J[\bar{K}^0 + \hat{\delta}K(t), \bar{L}^0 + \hat{\delta}L(t), \hat{\tau}] < \bar{J}^0, \quad (34)$$

where $\bar{J}^0$ is the cost corresponding to the optimal steady-state solution of the problem, using the static gains

$\bar{K}^0, \bar{L}^0$. Here, a weak variation in the controller gains from the optimal steady-state solution yields a lower cost.

Definition 14. Let $(\bar{P}, \bar{K}, \bar{L})$ be a steady-state admissible triple. The optimal periodic control problem is *normal* at $(\bar{P}, \bar{K}, \bar{L})$ if the following condition is satisfied for some τ:

$$\operatorname{rank}[(e^{\bar{F}\tau} - I_n)\,\bar{G}\;\bar{F}\bar{G}\;\cdots\;\bar{F}^{n-1}\bar{G}] = n. \tag{35}$$

Remark 15. Note that $(\bar{F}, \bar{G})$ controllability is sufficient to ensure normality.

For convenience, we will drop the use of functional notation for $\mathscr{H}$ and its derivatives—any usage is assumed to occur at the stationary point. Let us define $\mathbf{u}(s) \triangleq [\delta\mathbf{k}(s)^{\mathrm{T}}\ \delta\mathbf{l}(s)^{\mathrm{T}}]^{\mathrm{T}}$. Using the techniques of the previous subsections, we can construct $\mathscr{H}_{\mathbf{up}}, \mathscr{H}_{\mathbf{uu}}$, and $H(s)$, where $H(s)$ is the transfer function from $\mathbf{u}(s)$ to $\mathbf{p}(s)$. We also know that $\mathscr{H}_{\mathbf{pp}} = 0, \mathscr{H}_{\mathbf{pu}} = \mathscr{H}_{\mathbf{up}}^{\mathrm{T}}$.

Theorem 16. *If the local minimum of the optimal steady-state problem is normal and the $(m \times m)$ Hermitian matrix*

$$\Pi(\omega) = \mathscr{H}_{\mathbf{pu}}^{\mathrm{T}} H(j\omega) + H(-j\omega)^{\mathrm{T}}\mathscr{H}_{\mathbf{pu}} + \mathscr{H}_{\mathbf{uu}} \tag{36}$$

is partially negative for some $\omega > 0$, then the optimal periodic control problem is locally proper (and hence proper). Conversely, if the optimal periodic control problem is locally proper, then there exists $\omega > 0$ such that $\Pi(\omega)$ is not positive definite.

Proof. The proof for this theorem is the same as that given in Bittanti et al. (1973) and Bernstein and Gilbert (1980), where the control input is the vectorized parameters $\mathbf{u}$. □

Corollary 17 (Implications for strictly proper plants). *If the plant transfer function is strictly proper (i.e., $D = 0$), then there is a frequency Ω such that the optimal periodic control problem cannot be locally proper for frequencies greater than Ω.*

Proof. The magnitude of $H(j\omega)$ must attenuate at high frequencies due to the asymptotic stability of the stationary solution. Hence, $\Pi(j\omega) \to \mathscr{H}_{\mathbf{uu}}$ as $\omega \to \infty$. This means that if the optimization problem satisfies the Legendre–Clebsch condition, $\Pi(j\omega)$ must be positive definite for large enough ω. Now, the elements of $\mathscr{H}_{\mathbf{uu}}$ are given by

$$\mathscr{H}_{\mathbf{kk}} = 2R \otimes [P_1 - P_{12} - P_{12}^{\mathrm{T}} + P_2], \tag{37}$$

$$\mathscr{H}_{\mathbf{ll}} = 2\Lambda_2 \otimes [\Gamma_2\Gamma_2^{\mathrm{T}}], \tag{38}$$

$$\mathscr{H}_{\mathbf{kl}} = 2D^{\mathrm{T}} \overset{\mathrm{T}}{\otimes} [P_{13}^{\mathrm{T}}\Lambda_{13}] + 2D^{\mathrm{T}} \overset{\mathrm{T}}{\otimes} [P_3\Lambda_3] + 2D^{\mathrm{T}} \overset{\mathrm{T}}{\otimes} [P_{23}^{\mathrm{T}}\Lambda_{23}], \tag{39}$$

so since $D = 0$, $\mathscr{H}_{\mathbf{uu}}$ is positive definite if and only if both $\mathscr{H}_{\mathbf{kk}}$ and $\mathscr{H}_{\mathbf{ll}}$ are positive definite. We know that R, Λ_2, and $\Gamma_2\Gamma_2^{\mathrm{T}}$ are all positive definite. The quantity $[P_1 - P_{12} - P_{12}^{\mathrm{T}} + P_2]$ must be positive definite, since P is positive definite and $[P_1 - P_{12} - P_{12}^{\mathrm{T}} + P_2] = [I\ -I\ 0]P[I\ -I\ 0]^{\mathrm{T}}$. Hence $\mathscr{H}_{\mathbf{kk}}$ and $\mathscr{H}_{\mathbf{ll}}$ are the Kronecker products of positive definite matrices, which means they are positive definite themselves. So $\Pi(j\omega)$ converges to a positive definite matrix as $\omega \to \infty$, implying that there is a frequency Ω such that $\Pi(j\omega)$ is positive definite for all $\omega > \Omega$. By the results of the previous theorem, the optimal periodic control problem cannot be locally proper for frequencies $\omega > \Omega$. □

We thus have the interesting result that a chattering control that is a weak variation from the static optimum can never produce a better cost than the static optimum for any plant with a strictly proper transfer function.

6. Designing periodic optimal controllers

Once we have determined that only periodic controller gains can strongly stabilize the system, or if the Π test has established that periodic gains offer better performance than static ones, it remains to design these gains. We base our design methodology on standard optimal periodic control design practices, such as those in Speyer (1996).

Note first that choosing periodic gains makes the system matrices A_{cl} and B_{cl} periodic by Eq. (14). This in turn makes the solution to the Lyapunov equation (16) periodic, with a period that is the least common multiple of the periods of the elements of A_{cl} and B_{cl} (from the Lyapunov Lemma of Bittanti, Bolzern, & Colaneri, 1985).

Hence, if we specify a periodic functional form for the gains K and L, such as

$$K = K_0 + \sum_{k=1}^{N} K_{k1}\sin(k\omega t) + K_{k2}\cos(k\omega t),$$

$$L = L_0 + \sum_{k=1}^{N} L_{k1}\sin(k\omega t) + L_{k2}\cos(k\omega t),$$

then we can optimize cost (15) with respect to the parameters $\omega, \{K_{k1}, K_{k2}, L_{k1}, L_{k2}\}$, and the elements of $P(0)$. This optimization is subject to the constraints that $P(0)$ is a positive definite matrix and that P is periodic with period $\tau = 2\pi/\omega$. Alternatively, a periodic spline function (DeBoor, 1978) for the gains can be chosen, with the constraints

$$K(0) = K(\tau), \quad \frac{\mathrm{d}}{\mathrm{d}t}K\bigg|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}K\bigg|_{t=\tau},$$

$$L(0) = L(\tau), \quad \frac{\mathrm{d}}{\mathrm{d}t}L\bigg|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}L\bigg|_{t=\tau}$$

and the collocation points as optimization parameters. Again, the elements of $P(0)$ appear as optimization parameters and in the constraints that P is τ-periodic and

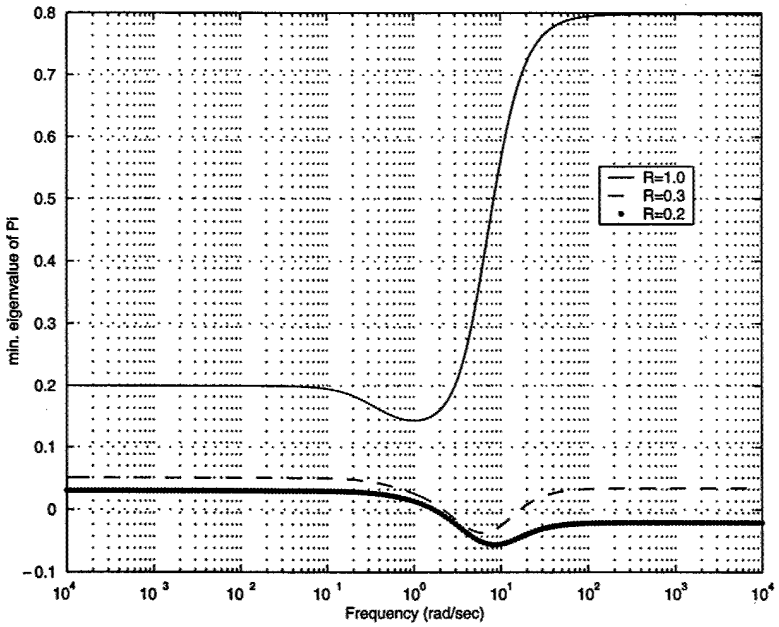


Fig. 1. Minimum eigenvalue of $\Pi(j\omega)$ vs. ω.

$P(0)$ is positive definite. The constraint that $P(0)$ is positive definite can be phrased in two ways:

(1) Require the leading principal minors of $P(0)$ to be positive.
(2) Parameterize $P(0)$ by its' $\mathcal{U}\mathcal{D}\mathcal{U}^{\mathrm{T}}$ decomposition $P(0) = \mathcal{U}(0)\mathcal{D}(0)\mathcal{U}(0)^{\mathrm{T}}$, where $\mathcal{U}$ is unit upper triangular and $\mathcal{D}$ is diagonal. $P(0)$ is then positive definite if and only if all diagonal elements of $\mathcal{D}(0)$ are positive. The periodicity constraints on P are transferred to the parameters $\mathcal{U}$ and $\mathcal{D}$, whose differential equations are given in Tapley and Peters (1980).

These nonlinearly constrained optimization problems can be solved by standard methods such as Sequential Quadratic Programming (Wilson, 1963; Boggs & Tolle, 1996) or accelerated gradient projection (Speyer, Kelley, Levine, & Denham, 1971).

Finally, note that there is no restriction in the above problem formulations requiring existence of a static solution. But nonlinear optimization problems should never be undertaken lightly—the Π test indicates when it is useful to attempt the difficult process of time-varying controller generation if a strongly stable LTI solution has already been found.

7. Examples

7.1. Π test for a plant with a DC term

Consider the linear system and cost given by

$$A = 1, \quad B = 1, \quad C = 1.5, \quad D = 1, \quad \Gamma_1 = 1,$$
$$\Gamma_2 = 1, \quad Q = 1, \quad R = 1, \quad Q_f = 0.01, \quad \Gamma_f = 1.$$

Note that the open loop transfer function, $(s+0.5)/(s-1)$, meets the parity interlacing property (Youla et al., 1974) and, therefore, the plant may be stabilized by a stable linear time invariant controller. However, the resulting conventional LQG controller is unstable.

A static solution for the modified cost given by (13) was found using the methods in Toivonen and Mäkilä (1985). The results of the local optimization were $K^0 = 3.9112$, $L^0 = 1.1774$. The pole of the static optimal controller was then -0.0724. The static optimum gains were also calculated for several other values of R. The Π test was then performed for each cost function and corresponding static optimal controller. For each case, the minimum eigenvalue of Π is plotted vs. frequency in Fig. 1. Note that when $R = 1$, the minimum eigenvalue of Π is never negative. Hence, there is no instance at which a lower cost can be realized via periodic gains. However, if R is reduced, the cost may potentially be reduced below the static optimum value. When $R = 0.3$, the minimum eigenvalue of Π falls below 0 for frequencies between 2 and 10.5 rad/s. If R is reduced to 0.2, the minimum eigenvalue of Π is negative for all frequencies greater than 2 rad/s, which means a chattering solution may reduce the cost below that of the static optimum. Note that the plant's transfer function is not strictly proper, so the restriction on the optimality of high frequency gains given by Corollary 17 does not apply.

7.2. Π test for a flexible structure

The problem of positioning the tip of a flexible robot arm using only sensors and actuators at the base of the arm

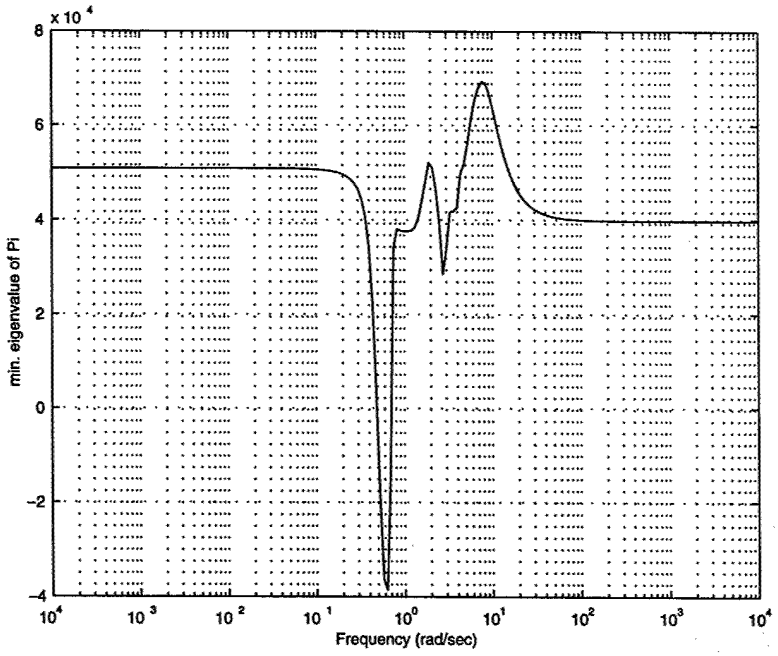


Fig. 2. Minimum eigenvalue of $\Pi(j\omega)$ vs. ω.

can be described by the following linear system and cost parameters:

$$A=\begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & -1 & 0 \end{bmatrix}, \quad B=\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix},$$

$$\Gamma_1=\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad Q=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$C=[1\ 0\ 0\ 0], \quad D=0, \quad \Gamma_2=1,$$

$$R=10^{-3}, \quad Q_f=10^{-2}I_4, \quad \Gamma_f=I_4,$$

where I_4 denotes the four-dimensional unit matrix. The open-loop transfer function for this system is $((s+j)(s-j))/(s^2(s+1.4142j)(s-1.4142j))$, which satisfies the parity interlacing property (Youla et al., 1974). This plant may thus be stabilized by a stable LTI controller. Despite this, the LQG gains yield an unstable controller.

A static solution for the modified cost given by (13) was found using the methods in Toivonen and Mäkilä (1985). The strongly stabilizing results of the local optimization were

$$K=[8.1188\ 2.0586\ -3.7766\ 4.9878],$$

$$L=[7.8756\ 6.0895\ 5.0344\ -1.7341]^{\mathrm{T}}.$$

The Π test was then performed; Fig. 2 plots the minimum eigenvalue of Π vs. frequency. Note that the high-frequency behavior is as Corollary 17 predicts, and that the optimization problem is locally proper only across a very narrow-frequency band.

7.3. Periodic optimal control

Strongly stabilizing periodic optimal controllers were generated for the one-dimensional plant described at the beginning of the section with parameters $Q=1$, $R=0.2$, using the methods described in Section 6. The performance of these controllers can be evaluated by comparing them to standard LQG optimal controllers, because the optimal LQG cost when the strong stability constraint is removed forms a lower bound for the cost that a strongly stabilizing controller can achieve. Using Sequential Quadratic Programming, we calculated strongly stabilizing optimal values for K and L when they were each parameterized by either three elements in a harmonic series or by 10 collocation points for a cubic spline spaced equally in a period. Table 1

Table 1
Efficiency comparison of compensators

Controller gain type	Cost by Eq. (9)	Cost by Eq. (13)
LQG	3.6544	Undefined
Static gains	4.0444	4.0992
Three harmonics	3.9636	4.0054
10 point spline	3.9295	3.9685

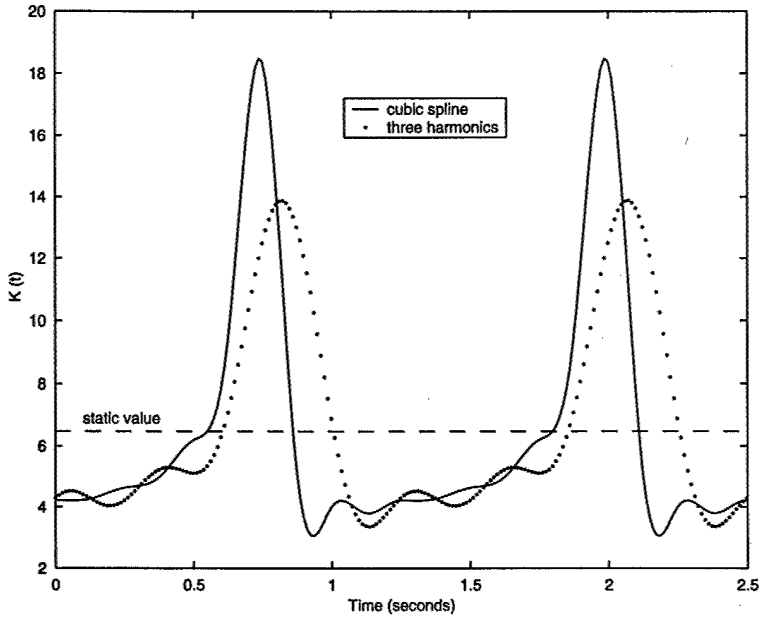


Fig. 3. Controller gain.

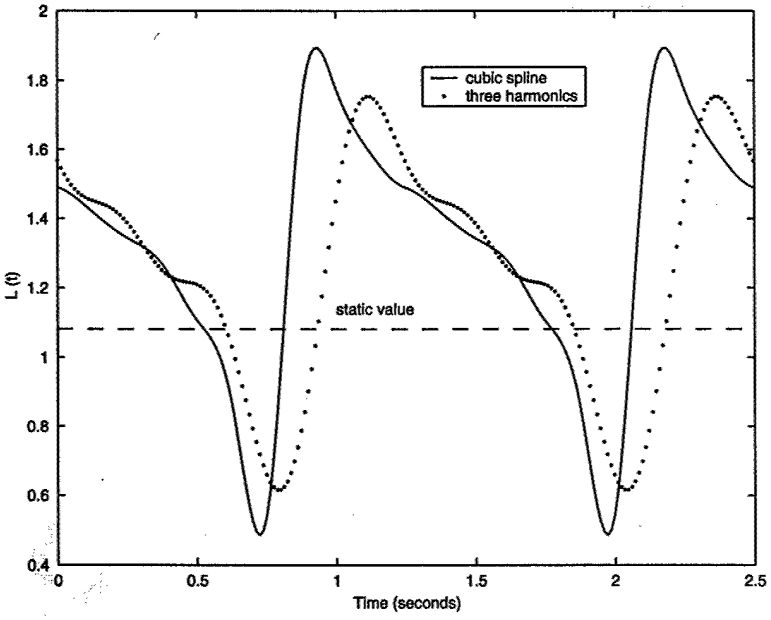


Fig. 4. Estimator gain.

compares the costs achieved by these compensators with the costs achieved by the strongly stabilizing static gains and the unstable LQG compensator. Note that the cost corresponding to the spline-parameterized strongly stabilizing controller was 29% closer to the lower bound LQG cost than that of the optimal strongly stabilizing static controller.

Figs. 3 and 4 plot the optimal values of the controller gain $K(t)$ and the estimator gain $L(t)$ over two periods, where the gains are parameterized both by a spline with 10 collocation points and by three harmonics of a Fourier series. Note that the shape of the gains are approximately the same for both parameterizations. More interestingly, observe that when the value of the controller gain is large, the value of the observer gain is small, and vice versa. The optimal strongly stabilizing periodic controllers thus oscillate between controller-dominant and observer-dominant phases in a smooth manner.

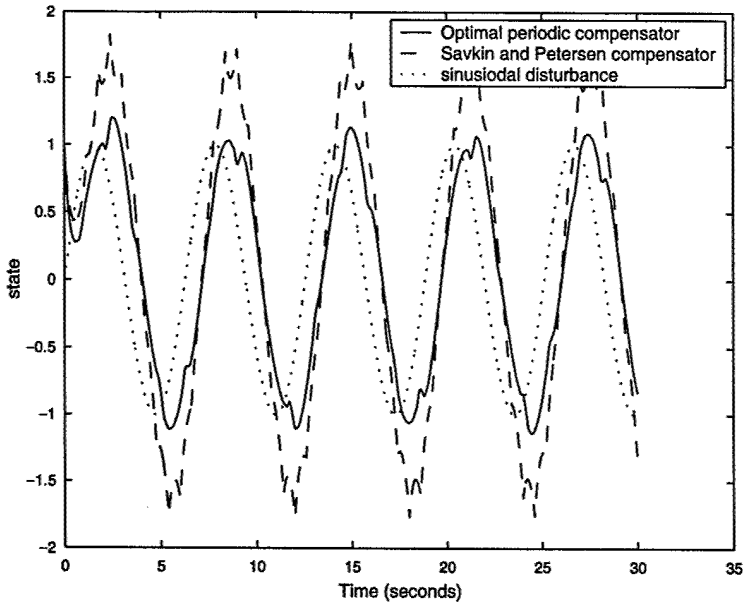


Fig. 5. Closed-loop responses to a sinusoidal disturbance.

The smoothness in the variation of the cubic spline controller's gains appears to enhance disturbance rejection relative to controllers constructed using the methods in Savkin and Petersen (1998). Fig. 5 illustrates the closed-loop responses to the disturbance $\sin(t)$ for both types of controller. Note also that the Savkin–Petersen controller for this example was chosen with the smallest possible sampling period that would guarantee strong stability. When larger sampling periods were used, the Savkin–Petersen closed-loop system exhibited even larger deviations.

8. Conclusion

A Π test applicable to a linear-quadratic-Gaussian strong stabilization problem has been developed, determining when periodic coefficients in the gain matrices can potentially reduce the cost. One important restriction to the test is that a stable, strictly proper controller of plant order must be found to ensure the existence of a strongly stabilizing static solution. Obviously, if no static solution exists, the optimal strongly stabilizing controller is time varying.

Techniques were then developed for synthesizing optimal periodic strongly stabilizing controllers. Because these techniques are computationally intensive, the Π test is valuable for determining in advance whether a periodic controller may improve performance. An example demonstrated that a strongly stable periodic optimal controller generated with our methods rejected persistent disturbances better than competing methods.

Methods used to derive the Π test in this can also be applied to other control problems. The material in Sections 3 and 4 enables many extensions to the work in Athans (1968) and Denham and Speyer (1964) on minimization of functions dependent on matrices. In particular, the techniques used here can be trivially modified to deal with problems involving optimizing decentralized controllers for systems with fixed modes (Wang & Davison, 1973).

Acknowledgements

This research was supported in part by the US Air Force Office of Scientific Research under Grant F49620-97-0272.

Appendix A. Constructing second partials of trace functions

The proofs of the assertions in Proposition 3 all follow the same general form. Therefore, only the construction of $(\partial^2/\partial y_{kl}\partial x_{ij})\,\mathbf{tr}(XAY^{\mathrm{T}}BYC)$ is provided here. Each of the other assertions may be constructed using similar arguments.

Let X have dimension $\bar{m} \times \bar{n}$ and let Y have dimension $\bar{o} \times \bar{p}$. We know from Athans (1968) that

$$\frac{\partial}{\partial x_{ij}} \mathbf{tr}(XAY^{\mathrm{T}}BYC)$$

$$= \frac{\partial}{\partial x_{ij}} \mathbf{tr}(AY^{\mathrm{T}}BYCX)$$

$$= [(AY^{\mathrm{T}}BYC)^{\mathrm{T}}]_{ij} = [C^{\mathrm{T}}Y^{\mathrm{T}}B^{\mathrm{T}}YA^{\mathrm{T}}]_{ij}. \quad \text{(A.1)}$$

Now, $[Y^T B^T Y]_{st}$ can be expressed as $\sum_{q=1}^{\bar{p}} \sum_{r=1}^{\bar{o}} y_{qs} b_{rq} y_{rt}$, so

$$[C^T Y^T B^T Y A^T]_{ij} = \sum_{s=1}^{\bar{p}} \sum_{t=1}^{\bar{m}} c_{si} [Y^T B^T Y]_{st} a_{jt}$$

$$= \sum_{s=1}^{\bar{p}} \sum_{t=1}^{\bar{m}} c_{si} \left(\sum_{q=1}^{\bar{p}} \sum_{r=1}^{\bar{o}} y_{qs} b_{rq} y_{rt} \right) a_{jt}. \quad \text{(A.2)}$$

Hence

$$\frac{\partial^2}{\partial y_{kl} \partial x_{ij}} \mathbf{tr}(XAY^T BYC) = \frac{\partial}{\partial y_{kl}} [C^T Y^T B^T Y A^T]_{ij}$$

$$= \sum_{t=1}^{\bar{m}} c_{li} \sum_{r=1}^{\bar{o}} b_{rk} y_{rt} a_{jt} + \sum_{s=1}^{\bar{p}} c_{si} \sum_{q=1}^{\bar{p}} y_{qs} b_{kq} a_{jl}$$

$$= c_{li} [B^T Y A^T]_{kj} + [BYC]_{ki} a_{jl}. \quad \text{(A.3)}$$

Appendix B. Second partials of the Hamiltonian

Using the definitions found in Section 5, the second partials of $\mathscr{H}$ have the following form:

$$\mathscr{H}_{KK} = 2R \otimes [P_1 - P_{12} - P_{12}^T + P_2], \quad \text{(B.1)}$$

$$\mathscr{H}_{LL} = 2\Lambda_2 \otimes [\Gamma_2 \Gamma_2^T], \quad \text{(B.2)}$$

$$\mathscr{H}_{KP_1} = [RK] \overset{T}{\otimes} I + [RK] \otimes I - [B^T \Lambda_1] \overset{T}{\otimes} I - [B^T \Lambda_1] \otimes I, \quad \text{(B.3)}$$

$$\mathscr{H}_{KP_2} = [B^T \Lambda_{12}] \otimes I + [B^T \Lambda_{12}] \overset{T}{\otimes} I + [RK] \otimes I + [RK] \overset{T}{\otimes} I, \quad \text{(B.4)}$$

$$\mathscr{H}_{KP_3} = -[B^T \Lambda_3] \otimes I - [B^T \Lambda_3] \overset{T}{\otimes} I + [D^T L^T \Lambda_3] \otimes I + [D^T L^T \Lambda_3] \overset{T}{\otimes} I, \quad \text{(B.5)}$$

$$\mathscr{H}_{KP_{12}} = -2[RK] \otimes I - 2[RK] \overset{T}{\otimes} I + 2[B^T \Lambda_1] \otimes I + 2[B^T \Lambda_{12}] \overset{T}{\otimes} I, \quad \text{(B.6)}$$

$$\mathscr{H}_{KP_{13}} = -2[B^T \Lambda_{13}^T] \otimes I + 2[D^T L^T \Lambda_{13}^T] \otimes I - 2[B^T \Lambda_{13}] \overset{T}{\otimes} I, \quad \text{(B.7)}$$

$$\mathscr{H}_{KP_{23}} = 2[B^T \Lambda_{13}] \overset{T}{\otimes} I - 2[B^T \Lambda_{23}^T] \otimes I + 2[D^T L^T \Lambda_{23}^T] \otimes I, \quad \text{(B.8)}$$

$$\mathscr{H}_{KL} = 2D^T \overset{T}{\otimes} [P_{13}^T \Lambda_{13}] + 2D^T \overset{T}{\otimes} [P_3 \Lambda_3] + 2D^T \overset{T}{\otimes} [P_{23}^T \Lambda_{23}], \quad \text{(B.9)}$$

$$\mathscr{H}_{LP_1} = 0, \quad \text{(B.10)}$$

$$\mathscr{H}_{LP_2} = -\Lambda_2 \overset{T}{\otimes} C - \Lambda_2 \otimes C, \quad \text{(B.11)}$$

$$\mathscr{H}_{LP_3} = -\Lambda_3 \overset{T}{\otimes} C + \Lambda_3 \overset{T}{\otimes} [DK] - \Lambda_3 \otimes C + \Lambda_3 \otimes [DK], \quad \text{(B.12)}$$

$$\mathscr{H}_{LP_{12}} = -2\Lambda_{12}^T \otimes C, \quad \text{(B.13)}$$

$$\mathscr{H}_{LP_{13}} = -2\Lambda_{13}^T \otimes C + 2\Lambda_{13}^T \otimes [DK], \quad \text{(B.14)}$$

$$\mathscr{H}_{LP_{23}} = -2\Lambda_{23}^T \otimes C + 2\Lambda_{23}^T \otimes [DK] - 2\Lambda_{23} \overset{T}{\otimes} C, \quad \text{(B.15)}$$

where I denotes a $n \times n$ identity matrix.

Appendix C. Linearized dynamics of the covariance

In Section 5, small variations were made to parameters in the covariance Lyapunov equation. Using the properties of the Kronecker and KT products, the expressions for the small variations in the covariance matrix can be written compactly as

$$\delta \dot{P}_1 = [(A - BK) \otimes I + I \otimes (A - BK)] \delta P_1 + [(BK) \overset{T}{\otimes} I + I \otimes (BK)] \delta P_{12} + [-B \otimes P_1 + B \otimes P_{12} - P_1 \overset{T}{\otimes} B + P_{12} \overset{T}{\otimes} B] \delta K, \quad \text{(C.1)}$$

$$\delta \dot{P}_2 = [(A - LC) \otimes I + I \otimes (A - LC)] \delta P_2 + [-I \otimes (P_2 C^T) - (P_2 C^T) \overset{T}{\otimes} I + I \otimes (L \Gamma_2 \Gamma_2^T) + (L \Gamma_2 \Gamma_2^T) \overset{T}{\otimes} I] \delta L, \quad \text{(C.2)}$$

$$\delta \dot{P}_3 = [(A - BK - LC + LDK) \otimes I + I \otimes (A - BK - LC + LDK)] \delta P_3 + [-B \otimes P_3 + (LD) \otimes P_3 - P_3 \overset{T}{\otimes} B + P_3 \overset{T}{\otimes} (LD)] \delta K + [-I \otimes (P_3 C^T) + I \otimes (P_3 K^T D^T) - (P_3 C^T) \overset{T}{\otimes} I + (P_3 K^T D^T) \overset{T}{\otimes} I] \delta L, \quad \text{(C.3)}$$

$$\delta \dot{P}_{12} = [(BK) \otimes I] \delta P_2 + [(A - BK) \otimes I + I \otimes (A - LC)] \delta P_{12} + [-B \otimes P_{12}^T + B \otimes P_2] \delta K + [-(P_{12} C^T) \overset{T}{\otimes} I] \delta L, \quad \text{(C.4)}$$

$$\delta \dot{P}_{13} = [(A - BK) \otimes I + I \otimes (A - BK - LC + LDK)] \times \delta P_{13} + [(BK) \otimes I] \delta P_{23} + [-B \otimes P_{13}^T$$

$$+B\otimes P_{23}^{\mathrm{T}}-P_{13}\overset{\mathrm{T}}{\otimes}B+P_{13}\overset{\mathrm{T}}{\otimes}(LD)]\delta K$$

$$+[-(P_{13}C^{\mathrm{T}})\overset{\mathrm{T}}{\otimes}I+(P_{13}K^{\mathrm{T}}D^{\mathrm{T}})\overset{\mathrm{T}}{\otimes}I]\delta L, \qquad \text{(C.5)}$$

$$\delta\dot{P}_{23}=[(A-LC)\otimes I+I\otimes(A-BK-LC+LDK)]\delta P_{23}$$

$$+[-P_{23}\overset{\mathrm{T}}{\otimes}B+P_{23}\overset{\mathrm{T}}{\otimes}(LD)]\delta K+[-I\otimes(P_{23}^{\mathrm{T}}C^{\mathrm{T}})$$

$$-(P_{23}C^{\mathrm{T}})\overset{\mathrm{T}}{\otimes}I+(P_{23}K^{\mathrm{T}}D^{\mathrm{T}})\overset{\mathrm{T}}{\otimes}I]\delta L. \qquad \text{(C.6)}$$

References

Athans, M. (1968). The matrix maximum principle. *Information and Control* (11), 592–606.

Bernstein, D. S., & Gilbert, E. G. (1980). Optimal periodic control: The π test revisited. *IEEE Transactions on Automatic Control, AC-25*(4), 673–684.

Bittanti, S., Bolzern, P., & Colaneri, P. (1985). The extended periodic lyapunov lemma. *Automatica, 21*(5), 603–605.

Bittanti, S., Fronza, G., & Guardabassi, G. (1973). Periodic control: A frequency domain approach. *IEEE Transactions on Automatic Control, AC-18*(1), 33–38.

Boggs, P. T., & Tolle, J. W. (1996). Sequential quadratic programming. *Acta Numerica* (4), 1–51.

De Boor, C. (1978). *A practical guide to splines*. New York: Springer.

Denham, W. F., & Speyer, J. L. (1964). Optimal measurement and velocity correction programs for midcourse guidance. *AIAA Journal, 2*(5), 896–907.

Geromel, J. C., & Bernussou, J. (1979). An algorithm for optimal decentralized regulation of linear quadratic interconnected systems. *Automatica, 15*, 489–491.

Lancaster, P. (1969). *Theory of matrices*. New York: Academic Press.

Savkin, A. V., & Petersen, I. R. (1998). Almost optimal LQ-control using stable periodic controllers. *Automatica, 34*(10), 1251–1254.

Speyer, J. L. (1996). Periodic optimal flight. *AIAA Journal of Guidance, Control, and Dynamics, 19*(4), 745–755.

Speyer, J. L., Kelley, H. J., Levine, N., & Denham, W. F. (1971). Accelerated gradient projection technique with application to rocket trajectory optimization. *Automatica, 7*(1), 37–43.

Tapley, B. D., & Peters, J. G. (1980). Sequential estimation algorithm using a continuous *udu*^{t} covariance factorization. *AIAA Journal of Guidance, Control, and Dynamics, 3*(4), 326–331.

Toivonen, H. T., & Mäkilä, P. M. (1985). A descent Anderson–Moore algorithm for optimal decentralized control. *Automatica, 21*(6), 743–744.

Toivonen, H. T., & Mäkilä, P. M. (1987). Newton's method for solving parametric linear quadratic control problems. *International Journal of Control, 46*(3), 897–911.

Vidyasagar, M. (1985). *Control system synthesis*. Cambridge, MA: MIT Press.

Wang, Y. W., & Bernstein, D. S. (1994). H_2-suboptimal stable stabilization. *Automatica, 30*(11), 1797–1800.

Wang, S. H., & Davison, E. D. (1973). On the stabilization of decentralized control systems. *IEEE Transactions on Automatic Control, AC-18*(5), 473–478.

Wilson, R. B. (1963). *A simplicial algorithm for concave programming*. Ph.D. thesis, Graduate School of Business Administration, Harvard University.

Youla, D. C., Bongiorno, J. J., & Lu, C. N. (1974). Single-loop feedback-stabilization of linear multivariable dynamical plants. *Automatica, 10*, 159–173.

Jonathan D. Wolfe received the B.S. and M.S. degrees in Mechanical Engineering from the University of California, Los Angeles in 1994, and the Ph.D. degree in Mechanical Engineering from same university in 2001. He is currently a Research Engineer in the Mechanical and Aerospace Engineering Department at the University of California, Los Angeles. His current research interests are differential GPS estimation and validation, fault-tolerant control, distributed data fusion, and multi-player games with information constraints.

Jason L. Speyer received the S.B. degree in aeronautics and astronautics from the Massachusetts Institute of Technology, Cambridge, in 1960 and the Ph.D. degree in applied mathematics from Harvard University, Cambridge, MA, in 1968.

His industrial experience includes research at Boeing, Raytheon, Analytical Mechanics Associated, and the Charles Stark Draper Laboratory. He was the Harry H. Power Professor in Aerospace Engineering at the University of Texas, Austin, and is currently a Professor in the Mechanical and Aerospace Engineering Department at the University of California, Los Angeles. He spent a research leave as a Lady Davis Visiting Professor at the Technion—Israel Institute of Technology, Haifa, Israel, in 1983 and was the 1990 Jerome C. Hunsaker Visiting Professor of Aeronautics and Astronautics at the Massachusetts Institute of Technology.

Dr. Speyer has twice been an elected member of the Board of Governors of the IEEE Control Systems Society. He has served as an Associate Editor of the *IEEE Transactions on Automatic Control* and as Chairman of the Technical Committee on Aerospace Control. He is a Fellow of the American Institute of Aeronautics and Astronautics and the Institute of Electrical and Electronic Engineers. From October 1987 to October 1991 and from October 1997 to October 2001, he has served as a member of the USAF Scientific Advisory Board. He was awarded Mechanics and Control of Flight Award and Dryden Lectureship in Research from the American Institute of Aeronautics and Astronautics in 1985 and 1995, respectively. He was awarded Air Force Exceptional Civilian Decoration in 1991 and 2001 and the IEEE Third Millennium Medal in 2000.

Appendix E

"A generalized least-squares fault detection filter,"

Robert H. Chen and Jason L. Speyer,

International Journal of Adaptive Control and Signal Processing, **vol. 14, pp. 747-757, 2000**

INTERNATIONAL JOURNAL OF ADAPTIVE CONTROL AND SIGNAL PROCESSING
Int. J. Adapt. Control Signal Process. 2000; **14**:747–757

A generalized least-squares fault detection filter

Robert H. Chen*,†, and Jason L. Speyer

Mechanical and Aerospace Engineering Department, University of California, Los Angeles, Los Angeles, CA 90095-1597, U.S.A.

SUMMARY

A fault detection and identification algorithm is determined from a generalization of the least-squares derivation of the Kalman filter. The objective of the filter is to monitor a single fault called the target fault and block other faults which are called nuisance faults. The filter is derived from solving a min–max problem with a generalized least-squares cost criterion which explicitly makes the residual sensitive to the target fault, but insensitive to the nuisance faults. It is shown that this filter approximates the properties of the classical fault detection filter such that in the limit where the weighting on the nuisance faults is zero, the generalized least-squares fault detection filter becomes equivalent to the unknown input observer where there exists a reduced-order filter. Filter designs can be obtained for both linear time-invariant and time-varying systems. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: fault detection and identification; unknown input observer; worst case design; time-varying system

1. INTRODUCTION

Any system under automatic control demands a high degree of system reliability. This requires a health monitoring system capable of detecting any plant, actuator and sensor fault as it occurs and identifying the faulty component. One approach, analytical redundancy which reduces the need for hardware redundancy, uses the modelled dynamic relationship between system inputs and measured system outputs to form a residual process used for detecting and identifying faults. A popular approach to analytical redundancy is the unknown input observer [1] which divides the faults into two groups: a single-target fault and possibly several nuisance faults. The nuisance faults are placed in an invariant subspace which is unobservable to the residual. Recently, approximate unknown input observers have been developed which have improved robustness to uncertainties and applicable to time-varying systems [2,3].

In this paper, a generalized least-squares fault detection filter, motivated by Chung and Speyer [2] and Bryson and Ho [4], is presented. A new least-squares problem with an indefinite cost

* Correspondence to: Robert H. Chen, Mechanical and Aerospace Engineering Department, University of California, Los Angeles, California 90095-1597, U.S.A.
† E-mail: chrobert@talus.seas.ucla.edu

Contract/grant sponsor: Air Force Office of Scientific Research; contract/grant number: F49620-97-1-0272
Contract/grant sponsor: California Department of Transportation; contract/grant number: 65A0013, MOU315

criterion is formulated as a min–max problem by generalizing the least-squares derivation of the Kalman filter [4] and allowing the explicit dependence on the target fault which is not presented in Reference [2]. Since the filter is derived similarly to Reference [2], many properties obtained in Reference [2] also apply to this filter. However, some new important properties are given. For example, since the target fault direction is now explicitly in the filter gain calculation, a mechanism is provided which enhances the sensitivity of the filter to the target fault. Furthermore, the projector, which annihilates the residual direction associated with the nuisance faults and is assumed in the problem formulation of Reference [2], is not required in the derivation of this filter. Finally, it is shown that this filter completely blocks the nuisance faults in the limit where the weighting on the nuisance faults is zero. For time-invariant systems, the nuisance faults are placed in a minimal (C, A)-unobservability subspace, and the generalized least-squares fault detection filter becomes equivalent to the unknown input observer. For time-varying systems, the nuisance faults are placed in a similar invariant subspace, and the generalized least-squares fault detection filter extends the unknown input observer to the time-varying case. In the limit, a reduced-order filter is derived for time-varying systems.

The problem is formulated in Section 2 and its solution is derived in Section 3 [2,4]. In Section 4, the filter is derived in the limit [2,5]. In Section 5, it is shown that, in the limit, the nuisance faults are placed in an invariant subspace. In Section 6, the reduced-order filter is derived in the limit. In Section 7, numerical examples are given.

2. PROBLEM FORMULATION

Consider a linear, observable system with two failure modes [1,2]

$$\dot{x} = Ax + Bu + F_1\mu_1 + F_2\mu_2 \tag{1a}$$

$$y = Cx + v \tag{1b}$$

where u is the control input, y is the measurement, v is the sensor noise, μ_1 is the target fault, and μ_2 is the nuisance fault. All system variables belong to real vector spaces, $x \in \mathscr{X}$, $u \in \mathscr{U}$, and $y \in \mathscr{Y}$. System matrices A, B, C, F_1 and F_2 are time-varying and continuously differentiable. The failure modes, μ_1 and μ_2, model the time-varying amplitude of the failure while the failure signatures, F_1 and F_2, model the directional characteristics of a failure. Assume F_1 and F_2 are monic so that $F_1 \neq 0$ and $F_2 \neq 0$ imply $F_1\mu_1 \neq 0$ and $F_2\mu_2 \neq 0$, respectively. In References [1,2], it is shown that this model, used to determine the fault detection filter, represents actuator, sensor and plant faults. There are two assumptions about the system (1) that are needed in order to have a well-conditioned unknown input observer. Assumption 2.1 ensures that the target fault can be isolated from the nuisance fault [1,2]. The output separability test is discussed in Remark 1 of Section 5. Assumption 2.2. ensures a non-zero residual in steady-state when the target fault occurs for time-invariant systems [3,6].

Assumption 2.1.

F_1 and F_2 are output separable.

Assumption 2.2.

For time-invariant systems, (C, A, F_1) does not have invariant zero at origin.

The objective of blocking the nuisance fault while detecting the target fault can be achieved by solving the following min–max problem:

$$\min_{\mu_1}\max_{\mu_2}\max_{x(t_0)}\frac{1}{2}\int_{t_0}^{t}(\|\mu_1\|^2_{Q_1^{-1}}-\|\mu_2\|^2_{\gamma Q_2^{-1}}-\|y-Cx\|^2_{V^{-1}})\,d\tau-\frac{1}{2}\|x(t_0)-\hat{x}_0\|^2_{\Pi_0} \tag{2}$$

subject to (1a). Note that, without the minimization with respect to μ_1, (2) reduces to the standard least-squares derivation of the Kalman filter [4]. t is the current time and y is assumed given. Q_1, Q_2, V and Π_0 are positive definite. γ is a non-negative scalar. Note that Q_1, Q_2, Π_0 and γ are design parameters to be chosen while V may be physically related to the power spectral density of the sensor noise because of (1b) [4]. The interpretation of the min–max problem is the following. Let μ_1^*, μ_2^* and $x^*(t_0)$ be the optimal strategies for μ_1, μ_2 and $x(t_0)$, respectively. Then, $x^*(\tau|Y_t)$, the x associated with μ_1^*, μ_2^* and $x^*(t_0)$, is the optimal trajectory for x where $\tau\in[t_0, t]$ and given the measurement history $Y_t=\{y(\tau)|t_0\leqslant\tau\leqslant t\}$. Since μ_1 maximizes $y-Cx$ and μ_2 minimizes $y-Cx$, $y-Cx^*$ is made primarily sensitive to μ_1 and minimally sensitive to μ_2. However, since x^* is the smoothed estimate of the state, a filtered estimate of the state, called $\hat{x}$, is needed for implementation. From the boundary condition in Section 3, at the current time t, $x^*(t|Y_t)=\hat{x}(t)$. Therefore, $y-C\hat{x}$ is primarily sensitive to the target fault and minimally sensitive to the nuisance fault. Note that when Q_1 is larger, $y-C\hat{x}$ is more sensitive to the target fault. When γ is smaller, $y-C\hat{x}$ is less sensitive to the nuisance fault. In Reference [2], the differential game blocks the nuisance fault, but does not enhance the sensitivity to the target fault. In Section 5, it is shown that the filter completely blocks the nuisance fault when γ is zero by placing it into an invariant subspace, called Ker S. Therefore, the residual used for detecting the target fault is

$$r=\hat{H}(y-C\hat{x}) \tag{3}$$

where $\hat{x}$, the filtered estimate of the state, is given in Section 3 and

$$\hat{H}:\mathscr{Y}\to\mathscr{Y},\qquad \operatorname{Ker}\hat{H}=C\operatorname{Ker}S,\quad \hat{H}=I-C\operatorname{Ker}S[(C\operatorname{Ker}S)^{\mathrm{T}}C\operatorname{Ker}S]^{-1}(C\operatorname{Ker}S)^{\mathrm{T}} \tag{4}$$

Ker S is given and discussed in Sections 4 and 5.

3. SOLUTION

In this section, the min–max problem given by (2) is solved [2,4]. The variational Hamiltonian of the problem is

$$\mathscr{H}=\tfrac{1}{2}(\|\mu_1\|^2_{Q_1^{-1}}-\|\mu_2\|^2_{\gamma Q_2^{-1}}-\|y-Cx\|^2_{V^{-1}})+\lambda^{\mathrm{T}}(Ax+Bu+F_1\mu_1+F_2\mu_2)$$

where $\lambda\in\mathscr{R}^n$ is a continuously differentiable Lagrange multiplier. The first-order necessary conditions [4] imply that the optimal strategies for μ_1, μ_2 and the dynamics for λ are

$$\mu_1^*=-Q_1F_1^{\mathrm{T}}\lambda,\quad \mu_2^*=\frac{1}{\gamma}Q_2F_2^{\mathrm{T}}\lambda,\quad \dot{\lambda}=-A^{\mathrm{T}}\lambda-C^{\mathrm{T}}V^{-1}(y-Cx)$$

with boundary conditions

$$\lambda(t_0) = \Pi_0[x^*(t_0) - \hat{x}_0], \quad \lambda(t) = 0 \tag{5}$$

By substituting μ_1^* and μ_2^* into (1a), the two-point boundary value problem requires the solution to

$$\begin{bmatrix} \dot{x}^* \\ \dot{\lambda} \end{bmatrix} = \begin{bmatrix} A & \frac{1}{\gamma}F_2Q_2F_2^{\mathrm{T}} - F_1Q_1F_1^{\mathrm{T}} \\ C^{\mathrm{T}}V^{-1}C & -A^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} x^* \\ \lambda \end{bmatrix} + \begin{bmatrix} Bu \\ -C^{\mathrm{T}}V^{-1}y \end{bmatrix} \tag{6}$$

with boundary conditions (5). The form of (5) suggests that

$$\lambda = \Pi(x^* - \hat{x}) \tag{7}$$

where $\Pi(t_0) = \Pi_0$, $\hat{x}(t_0) = \hat{x}_0$ and $\hat{x}$ is an intermediate state. By differentiating (7), using (6), adding and subtracting $\Pi A\hat{x}$ and $C^{\mathrm{T}}V^{-1}C\hat{x}$, the following dynamic filter structure results:

$$\Pi\dot{\hat{x}} = \Pi A\hat{x} + \Pi Bu + C^{\mathrm{T}}V^{-1}(y - C\hat{x}), \quad \hat{x}(t_0) = \hat{x}_0 \tag{8}$$

$$-\dot{\Pi} = \Pi A + A^{\mathrm{T}}\Pi + \Pi\left(\frac{1}{\gamma}F_2Q_2F_2^{\mathrm{T}} - F_1Q_1F_1^{\mathrm{T}}\right)\Pi - C^{\mathrm{T}}V^{-1}C, \quad \Pi(t_0) = \Pi_0 \tag{9}$$

Since $x^* = \hat{x}$ at current time t (5), the generalized least-squares fault detection filter is (8). Note that (8) is used by the residual (3) to detect the target fault.

4. LIMITING CASE

In this section, the min–max problem (2) is solved in the limit where γ is zero [2,5]. When γ is zero, there is no constraint on μ_2 to minimize $y - Cx$. Therefore, the nuisance fault is completely blocked from the residual which is shown in Section 5.

In the limit, the min–max problem (2) becomes

$$\min_{\mu_1}\max_{\mu_2}\max_{x(t_0)}\frac{1}{2}\int_{t_0}^{t}(\|\mu_1\|^2_{Q_1^{-1}} - \|y - Cx\|^2_{V^{-1}})\,\mathrm{d}\tau - \frac{1}{2}\|x(t_0) - \hat{x}_0\|^2_{\Pi_0} \tag{10}$$

This problem is singular with respect to μ_2. Therefore, the Goh transformation [5] is used to form a non-singular problem. Let

$$\phi_1(\tau) = \int_{t_0}^{\tau}\mu_2(s)\,\mathrm{d}s, \quad \alpha_1 = x - F_2\phi_1$$

By differentiating α_1 and using (1a),

$$\dot{\alpha}_1 = A\alpha_1 + Bu + F_1\mu_1 + B_1\phi_1 \tag{11}$$

where $B_1 = AF_2 - \dot{F}_2$. By substituting α_1 into (10), the new min–max problem is

$$\min_{\mu_1}\max_{\phi_1}\max_{\alpha_1(t_0^+)}\frac{1}{2}\int_{t_0}^{t}[\|\mu_1\|^2_{Q_1^{-1}} - \|\phi_1\|^2_{F_2^TC^TV^{-1}CF_2} - \|y - C\alpha_1\|^2_{V^{-1}} + (y - C\alpha_1)^TV^{-1}CF_2\phi_1$$

$$+ \phi_1^TF_2^TC^TV^{-1}(y - C\alpha_1)]\,d\tau - \frac{1}{2}\|\alpha_1(t_0^+) + F_2\phi_1(t_0^+) - \hat{x}_0\|^2_{\Pi_0} \tag{12}$$

subject to (11). If $F_2^TC^TV^{-1}CF_2$ fails to be positive definite, (12) is still a singular problem with respect to ϕ_1. Then, the Goh transformation has to be used until the problem becomes non-singular. If $F_2^TC^TV^{-1}CF_2 = 0$, let

$$\phi_2(\tau) = \int_{t_0}^{\tau}\phi_1(s)\,ds, \quad \alpha_2 = \alpha_1 - B_1\phi_2$$

Then, $\dot{\alpha}_2 = A\alpha_2 + Bu + F_1\mu_1 + B_2\phi_2$ where $B_2 = AB_1 - \dot{B}_1$. If $F_2^TC^TV^{-1}CF_2 \geqslant 0$, the Goh transformation is applied only on the singular part [6]. The transformation process stops if the weighting on ϕ_2, $B_1^TC^TV^{-1}CB_1$, is positive definite. Otherwise, continue the transformation until there exists B_k such that the weighting on ϕ_k, $B_{k-1}^TC^TV^{-1}CB_{k-1}$, is positive definite. Then, in the limit, the min–max problem (2) becomes

$$\min_{\mu_1}\max_{\phi_k}\max_{\alpha_k(t_0^+)}\frac{1}{2}\int_{t_0}^{t}[\|\mu_1\|^2_{Q_1^{-1}} - \|\phi_k\|^2_{B_{k-1}^TC^TV^{-1}CB_{k-1}} - \|y - C\alpha_k\|^2_{V^{-1}} + (y - C\alpha_k)^TV^{-1}CB_{k-1}\phi_k$$

$$+ \phi_k^TB_{k-1}^TC^TV^{-1}(y - C\alpha_k)]\,d\tau - \frac{1}{2}\|\alpha_k(t_0^+) + \bar{B}\bar{\phi}(t_0^+) - \hat{x}_0\|^2_{\Pi_0} \tag{13}$$

subject to $\dot{\alpha}_k = A\alpha_k + Bu + F_1\mu_1 + B_k\phi_k$ where $\bar{B} = [F_2\ B_1\ B_2\ \cdots\ B_{k-1}]$ and $\bar{\phi} = [\phi_1^T\ \phi_2^T\ \cdots\ \phi_k^T]^T$. The min–max problem (13) can be solved similarly to (2). Therefore, the derivation [6] is not repeated here. The limiting generalized least-squares fault detection filter is

$$S\dot{\hat{x}} = SA\hat{x} + SBu + [SB_k(B_{k-1}^TC^TV^{-1}CB_{k-1})^{-1}B_{k-1}^TC^TV^{-1} + C^T\bar{H}^TV^{-1}\bar{H}](y - C\hat{x}) \tag{14}$$

where

$$-\dot{S} = S\bar{A} + \bar{A}^TS + S[B_k(B_{k-1}^TC^TV^{-1}CB_{k-1})^{-1}B_k^T - F_1Q_1F_1^T]S - C^T\bar{H}^TV^{-1}\bar{H}C \tag{15}$$

$\bar{H} = I - CB_{k-1}(B_{k-1}^TC^TV^{-1}CB_{k-1})^{-1}B_{k-1}^TC^TV^{-1}$ and $\bar{A} = A - B_k(B_{k-1}^TC^TV^{-1}CB_{k-1})^{-1}B_{k-1}^T$ $C^TV^{-1}C$ subject to $\hat{x}(t_0^+) = \hat{x}_0$ and $S(t_0^+) = \Pi_0 - \Pi_0\bar{B}(\bar{B}^T\Pi_0\bar{B})^{-1}\bar{B}^T\Pi_0$. However, (14) cannot be used because S has a null space which is shown in Theorem 4.1. Therefore, a reduced-order filter for (14) is derived in Section 6.

Theorem 4.1.

$$S[B_{k-1}\ B_{k-2}\ \cdots\ B_1\ F_2] = 0.$$

Proof. The proof is similar to Reference [2] and can be found in Reference [6]. □

5. PROPERTIES OF THE NULL SPACE OF S

In this section, some properties of the null space of S are given. It is shown that the null space of S is equivalent to the minimal (C, A) – unobservability subspace for time-invariant systems and a similar invariant subspace for time-varying systems. Therefore, the limiting generalized least-squares fault detection filter is equivalent to the unknown input observer and extends it to the time-varying case. The minimal (C, A)-unobservability subspace is a subspace which is $(A - LC)$-invariant and unobservable with respect to $(\tilde{H}C, A - LC)$ for some filter gain L and projector $\tilde{H}$ [1]. One method for computing the minimal (C, A)-unobservability subspace of F_2, called $\mathscr{T}_2$ here, is $\mathscr{T}_2 = \mathscr{W}_2 \oplus \mathscr{V}_2$ [1] where $\mathscr{W}_2 = [B_{k-1}\ B_{k-2} \cdots B_1\ F_2]$ is the minimal (C, A)-invariant subspace of F_2 and $\mathscr{V}_2$ is the subspace spanned by the invariant zero directions of (C, A, F_2). Note that the associated $\tilde{H}$ is

$$\tilde{H}: \mathscr{Y} \to \mathscr{Y}, \quad \text{Ker } \tilde{H} = CB_{k-1}, \quad \tilde{H} = I - CB_{k-1}[(CB_{k-1})^{\mathrm{T}} CB_{k-1}]^{-1}(CB_{k-1})^{\mathrm{T}} \tag{16}$$

Note that Ker $\tilde{H}$ = Ker $\bar{H}$.

Theorem 5.1 shows that the null space of S is a (C, A)-invariant subspace. Theorem 5.2 shows that the null space of S is contained in the unobservable subspace of $(\tilde{H}C, A - LC)$.

Theorem 5.1.

Ker S is a (C, A)-invariant subspace.

Proof. The dynamic equation of the error, $e = x - \hat{x}$, in the absence of the target fault and sensor noise can be obtained by using (1) and (14):

$$S\dot{e} = [SA + SB_k(B_{k-1}^{\mathrm{T}} C^{\mathrm{T}} V^{-1} CB_{k-1})^{-1} B_{k-1}^{\mathrm{T}} C^{\mathrm{T}} V^{-1} C + C^{\mathrm{T}} \bar{H}^{\mathrm{T}} V^{-1} \bar{H} C]e$$

because $SF_2 = 0$. By adding $\dot{S}e$ to both sides and using (15),

$$\frac{\mathrm{d}}{\mathrm{d}\tau}(Se) = -\{[A - B_k(B_{k-1}^{\mathrm{T}} C^{\mathrm{T}} V^{-1} CB_{k-1})^{-1} B_{k-1}^{\mathrm{T}} C^{\mathrm{T}} V^{-1} C]^{\mathrm{T}}$$

$$+ S[-F_1 Q_1 F_1^{\mathrm{T}} + B_k(B_{k-1}^{\mathrm{T}} C^{\mathrm{T}} V^{-1} CB_{k-1})^{-1} B_k^{\mathrm{T}}]\}Se \tag{17}$$

If the error initially lies in Ker S, (17) implies that the error will never leave Ker S. Therefore, Ker S is a (C, A)-invariant subspace. □

Theorem 5.2.

Ker S is contained in the unobservable subspace of $(\tilde{H}C, A - LC)$.

Proof. Let $\zeta \in$ Ker S. By multiplying (15) by ζ^{T} from the left and ζ from the right,

$$\frac{\mathrm{d}}{\mathrm{d}\tau}(\zeta^{\mathrm{T}} S\zeta) = \zeta^{\mathrm{T}} C^{\mathrm{T}} \bar{H}^{\mathrm{T}} V^{-1} \bar{H} C\zeta = 0$$

Then, $\tilde{H}C\zeta = 0$ because $\bar{H}C\zeta = 0$ and $\operatorname{Ker}\bar{H} = \operatorname{Ker}\tilde{H}$. From Theorem 5.1, $\operatorname{Ker} S$ is a (C, A)-invariant subspace. Therefore, $\operatorname{Ker} S$ is contained in the unobservable subspace of $(\tilde{H}C, A - LC)$. □

From Theorem 4.1, $C \operatorname{Ker} S \supseteq CB_{k-1}$. From Theorem 5.2, $C \operatorname{Ker} S \subseteq CB_{k-1}$. Therefore, $C \operatorname{Ker} S = CB_{k-1}$ and $\hat{H}$ (4) is equivalent to $\tilde{H}$ (16). Note that (16) is a better way to form $\hat{H}$ which is used by the residual (3) because it does not require the solution to the limiting Riccati Equation (15).

For time-invariant systems, it is important to discuss the invariant zero directions when designing the fault detection filter. The invariant zeros of (C, A, F_2) will become part of the eigenvalues of the filter if their associated invariant zero directions are not included in the invariant subspace of F_2 [1]. From Reference [3,6], the null space of S includes all the invariant zero directions if the nuisance fault direction is modified to the invariant zero directions. Therefore, the invariant zeros will not become part of the filter eigenvalues. From Theorem 4.1 and modified nuisance fault direction, the null space of S contains the minimal (C, A)-unobservability subspace of F_2. By combining with Theorem 5.2, the null space of S is equivalent to the minimal (C, A)-unobservability subspace of F_2, and the limiting generalized least-squares fault detection filter is equivalent to the unknown input observer. Note that the invariant zero and minimal (C, A)-unobservability subspace are only defined for time-invariant systems. For time-varying systems, Theorems 4.1, 5.1 and 5.2 imply that the null space of S is a similar invariant subspace.

Remark 1.

In order to detect the target fault, F_1 cannot intersect the null space of S which is unobservable to the residual. If it does, the target fault will be difficult or impossible to detect even though the filter can still be derived by solving the min–max problem. If F_1 does not intersect the null space of S, F_1 and F_2 are called output separable [1], and the output separability test can be stated as $CB_{k-1} \cap C\tilde{B}_{k-1} = \emptyset$ where $\tilde{B}_{k-1}$ is the Goh transformation of F_1.

6. REDUCED-ORDER FILTER

In this section, the reduced-order filter is derived for the limiting generalized least-squares fault detection filter (14). The reduced-order filter is necessary for implementation because (14) cannot be used due to the null space of S. Since S is non-negative definite, there exists a state transformation Γ such that

$$\Gamma^{\mathrm{T}} S \Gamma = \begin{bmatrix} \bar{S} & 0 \\ 0 & 0 \end{bmatrix} \tag{18}$$

where $\bar{S}$ is positive definite. Theorem 6.1 provides a way to form the transformation.

Theorem 6.1.

There exists a state transformation Γ where

$$[Z \ \operatorname{Ker} S] = \Gamma \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \tag{19}$$

Z is any $n \times (n - k_2)$ continuously differentiable matrix such that itself and Ker S span the state space where $n = \dim \mathscr{X}$ and $k_2 = \dim(\text{Ker } S)$. Z_1 and Z_2 are any $(n - k_2) \times (n - k_2)$ and $k_2 \times k_2$ invertible continuously differentiable matrices, respectively. Then, the Γ obtained from (19) satisfies (18).

Proof.

$$\text{Ker } S = \Gamma \begin{bmatrix} 0 \\ Z_2 \end{bmatrix} \Rightarrow S\Gamma \begin{bmatrix} 0 \\ Z_2 \end{bmatrix} = 0 \Rightarrow \Gamma^{\mathrm{T}} S \Gamma \begin{bmatrix} 0 \\ Z_2 \end{bmatrix} = 0$$

Since Z_2 is invertible by definition and $\Gamma^{\mathrm{T}} S \Gamma$ is symmetric, (18) is true. □

Note that Theorem 6.1 does not define Γ uniquely and Γ can be computed *a priori* because Ker S can be obtained *a priori*.

By applying the transformation to the estimator state, $\Gamma^{-1}\hat{x} \triangleq \hat{\eta} = [\hat{\eta}_1^{\mathrm{T}} \ \hat{\eta}_2^{\mathrm{T}}]^{\mathrm{T}}$. By multiplying (14) by Γ^{T} from the left, using $\Gamma\Gamma^{-1} = I$, and adding $\Gamma^{\mathrm{T}} S \Gamma \dot{\Gamma}^{-1} \hat{x}$ to both sides, the limiting filter can be transformed into two equations,

$$\bar{S}\dot{\hat{\eta}}_1 = \bar{S}(A_{11} - \Gamma_{11})\hat{\eta}_1 + \bar{S}(A_{12} - \Gamma_{12})\hat{\eta}_2 + \bar{S}M_1 u + [\bar{S}G_1(D_2^{\mathrm{T}}C_2^{\mathrm{T}}V^{-1}C_2D_2)^{-1}D_2^{\mathrm{T}}C_2^{\mathrm{T}}V^{-1} + C_1^{\mathrm{T}}\bar{H}^{\mathrm{T}}V^{-1}\bar{H}](y - C_1\hat{\eta}_1 - C_2\hat{\eta}_2) \tag{20a}$$

$$0 = C_2^{\mathrm{T}}\bar{H}^{\mathrm{T}}V^{-1}\bar{H}(y - C_1\hat{\eta}_1 - C_2\hat{\eta}_2) \tag{20b}$$

where

$$\Gamma^{-1}\dot{\Gamma} = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix}, \quad \Gamma^{-1}A\Gamma = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad \Gamma^{-1}B = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}, \quad C\Gamma = [C_1 \ C_2]$$

$$\Gamma^{-1}F_1 = \begin{bmatrix} N_1 \\ N_2 \end{bmatrix}, \quad \Gamma^{-1}B_{k-1} = \begin{bmatrix} 0 \\ D_2 \end{bmatrix}, \quad \Gamma^{-1}B_k = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}$$

Note that Γ^{-1} and $\dot{\Gamma}$ can be computed *a priori* from (19). From (20b),

$$\bar{H}C_2 = 0 \tag{21}$$

because $y - C_1\hat{\eta}_1 - C_2\hat{\eta}_2$ is arbitrary. By multiplying (15) by Γ^{T} from the left and Γ from the right, subtracting $\dot{\Gamma}^{\mathrm{T}} S \Gamma$ and $\Gamma S \dot{\Gamma}^{\mathrm{T}}$ from both sides, and using $\Gamma\Gamma^{-1} = I$, the limiting Riccati equation can be transformed into two equations,

$$0 = \bar{S}[A_{12} - \Gamma_{12} - G_1(D_2^{\mathrm{T}}C_2^{\mathrm{T}}V^{-1}C_2D_2)^{-1}D_2^{\mathrm{T}}C_2^{\mathrm{T}}V^{-1}C_2] \tag{22}$$

Note that this filter is equivalent to the optimal stochastic fault detection filter [12] which is an approximate unknown input observer.

4. LIMITING CASE

In this section, the robust multiple-fault detection filter is determined in the limit as $\gamma_i \to 0$, $i = 1 \cdots s$, when there is no complementary subspace. It is shown that, if $s = q$, the filter places each associated nuisance fault into the unobservable subspace of its associated projected residual for both time-invariant and time-varying systems. Therefore, the filter becomes equivalent to the RDD filter in the limit and extends the RDD filter to the time-varying case. In Section 4.1, the geometric structure of the detection filter is given [3]. In Section 4.2, the robust multiple-fault detection filter is determined in the limit. In Section 4.3, the conditions to ensure that the faults can be isolated are discussed.

4.1. Geometric structure of detection filter

The BJD filter places each fault μ_i into an invariant subspace $\mathscr{T}_i$ [3] where

$$\mathscr{T}_i = \mathscr{W}_i \oplus \mathscr{V}_i \tag{21}$$

$\mathscr{T}_i$ is called the minimal (C,A)-unobservability subspace or the detection space of F_i. $\mathscr{W}_i$ is the minimal (C,A)-invariant subspace of F_i given by

$$\mathscr{W}_i = \mathrm{Im}[f_{i,1} \ \cdots \ A^{\delta_{i,1}} f_{i,1} \ f_{i,2} \ \cdots \ A^{\delta_{i,2}} f_{i,2} \ \cdots \ f_{i,p_i} \ \cdots \ A^{\delta_{i,p_i}} f_{i,p_i}] \tag{22}$$

where $f_{i,j}$ is the jth column of F_i, $\delta_{i,j}$ is the smallest non-negative integer such that $CA^{\delta_{i,j}} f_{i,j} \neq 0$ and $p_i = \dim F_i$. $\mathscr{V}_i$ is the subspace spanned by the invariant zero directions of (C,A,F_i). The RDD filter places each associated nuisance fault $\hat{\mu}_i$ into an invariant subspace $\hat{\mathscr{T}}_i = [\mathscr{T}_1 \cdots \mathscr{T}_{i-1} \ \mathscr{T}_{i+1} \cdots \mathscr{T}_q]$ which is the unobservable subspace of $(\hat{H}_i C, A - LC)$ where L is the filter gain and $\hat{H}_i$ is given in (9) [3]. Therefore, each associated nuisance fault is in the unobservable subspace of its associated projected residual.

For time-varying systems, the minimal (C,A)-invariant subspace of F_i is [10]

$$\mathscr{W}_i = \mathrm{Im}[b_{i,1,0} \ \cdots \ b_{i,1,\delta_{i,1}} \ b_{i,2,0} \ \cdots \ b_{i,2,\delta_{i,2}} \ \cdots \ b_{i,p_i,0} \ \cdots \ b_{i,p_i,\delta_{i,p_i}}] \tag{23}$$

which is found from the iteration defined by the Goh transformation (10). For time-varying systems, the minimal (C,A)-unobservability subspace cannot be determined by (21) because the concept of invariant zero is for time-invariant systems only.

Remark 1

Equations (22) and (23) produce the correct invariant subspaces only when $\mathrm{Rank}(C\mathscr{W}_i) = p_i$. If $\mathrm{Rank}(C\mathscr{W}_i) < p_i$, a new basis for F_i can be obtained such that $\mathrm{Rank}(C\mathscr{W}_i) = p_i$ [17]. For example, for time-invariant systems, $F_i = [f_{i,1} \ f_{i,2}]$ where $f_{i,1} \neq f_{i,2}$ and $Cf_{i,1} = Cf_{i,2} \neq 0$. Then, $\mathscr{W}_i = \mathrm{Im}[f_{i,1} \ f_{i,2}]$ from (22). Since $\mathrm{Rank}(C\mathscr{W}_i) = 1$, (22) does not produce the correct invariant subspace. By using a different basis for F_i, e.g. $[f_{i,1} \ f_{i,1} - f_{i,2}]$, $\mathscr{W}_i = \mathrm{Im}[f_{i,1} \ f_{i,1} - f_{i,2} \ A(f_{i,1} - f_{i,2})]$ from (22) which is equivalent to $\mathrm{Im}[f_{i,1} \ f_{i,2} \ A(f_{i,1} - f_{i,2})]$. Since $\mathrm{Rank}(C\mathscr{W}_i) = 2$, (22) produces the correct invariant subspace using this new basis of F_i. This invariant subspace can also be confirmed by using the recursive algorithm given in Reference [3].

4.2. *Limiting robust multiple-fault detection filter*

In this section, the robust multiple-fault detection filter is determined in the limit as $\gamma_i \to 0$, $i = 1 \cdots s$, when there is no complementary subspace and $s = q$. The filter for time-invariant systems is considered first. Then, the filter for time-varying systems is considered in Remark 3 at the end of this section. First, it is assumed that in the limit, $\hat{\mathscr{T}}_1 \cdots \hat{\mathscr{T}}_q$ are $(A - LC)$-invariant where L is in (14). This will be shown to be true later. Then, the filter gain (14) is simplified in the limit by using Lemma 4.1 so that the simplified filter gain does not require the solution to the two-point boundary value problem, (12) and (15). Lemma 4.2 shows that the simplified filter gain minimizes the cost criterion. Therefore, the simplified filter gain is equivalent to (14) in the limit. Lemma 4.2 also shows that (9) is the optimal projector in the limit. Finally, Theorem 4.3 shows that $\hat{\mathscr{T}}_1 \cdots \hat{\mathscr{T}}_q$ are $(A - LC)$-invariant where L is the simplified filter gain. Therefore, the filter becomes equivalent to the RDD filter in the limit. Corollary 4.4 shows that $\mathscr{T}_1 \cdots \mathscr{T}_q$ are $(A - LC)$-invariant where L is the simplified filter gain. Therefore, the filter also becomes equivalent to the BJD filter in the limit.

Lemma 4.1
Define a new projector H_i where

$$H_i : \mathscr{X} \to \mathscr{X}, \quad \operatorname{Ker} H_i = \hat{\mathscr{T}}_i, \quad H_i = I - \hat{\mathscr{T}}_i (\hat{\mathscr{T}}_i^T \hat{\mathscr{T}}_i)^{-1} \hat{\mathscr{T}}_i^T$$

for $i = 1 \cdots q$. In the limit, H_i has the following properties:

$$\left(\sum_{j=1}^{q} H_j \right)^{-1} H_i = \left(\sum_{j=1}^{q} K_j \right)^{-1} K_i \tag{24a}$$

$$H_i W_i = 0 \tag{24b}$$

Proof
See Appendix A.1. □

In the limit, by applying Lemma 4.1 to (14),

$$L^* = \left(\sum_{i=1}^{q} H_i \right)^{-1} \left(\sum_{i=1}^{q} H_i P_i \right) C^T V^{-1} \tag{25}$$

Note that (25) does not require the solution to the two-point boundary value problem, (12) and (15), but just the solution to the Riccati equation (13) which can be obtained independently of L. By using the asymptotic expansion of P_i in Reference [12], it can be shown that $H_i P_i$ remains finite in the limit even though P_i goes to infinity. Therefore, the limiting filter gain (25) remains finite. Lemma 4.2 shows that (25) minimizes the cost criterion. Therefore, (25) is equivalent to (14) in the limit. Lemma 4.2 also shows that (9) is the optimal projector in the limit.

Lemma 4.2
In the limit, the cost criterion associated with (25) is zero.

Proof
See Appendix A.2. □

Remark 2
For the single-fault filter, the filter gain (20) goes to infinity in the limit and there exists a reduced-order filter [12]. For the multiple-fault filter, however, the limiting filter gain (25) remains finite.

Theorem 4.3 shows that $\hat{\mathscr{T}}_1 \cdots \hat{\mathscr{T}}_q$ are $(A - LC)$-invariant where L is in (25). Therefore, the filter becomes equivalent to the RDD filter in the limit. Corollary 4.4 shows that $\mathscr{T}_1 \cdots \mathscr{T}_q$ are $(A - LC)$-invariant where L is in (25). Therefore, the filter also becomes equivalent to the BJD filter in the limit.

Theorem 4.3
In the limit, $\hat{\mathscr{T}}_1 \cdots \hat{\mathscr{T}}_q$ are $(A - LC)$-invariant where L is in (25).

Proof
See Appendix A.3. □

Corollary 4.4
In the limit, $\mathscr{T}_1 \cdots \mathscr{T}_q$ are $(A - LC)$-invariant where L is in (25).

Proof
See Appendix A.4. □

Remark 3
For time-varying systems, the minimal (C, A)-unobservability subspace cannot be determined by (21) because the concept of invariant zero is for time-invariant systems only. However, by letting $\hat{\mathscr{T}}_i = \operatorname{Ker} \bar{\Pi}_i$ which is given in Appendix A.3, it can be shown that $\operatorname{Ker} \bar{\Pi}_i$ is included in the unobservable subspace of $(\hat{H}_i C, A - LC)$ where L is in (25) and $\hat{H}_i$ is given in (9) [11,12]. Furthermore, $\operatorname{Ker} \bar{\Pi}_i$ is equivalent to the unobservable subspace of $(\hat{H}_i C, A - LC)$ when there is no complementary subspace. Then, all the lemmas, theorem and corollary in this section can be shown similarly for time-varying systems. Therefore, the filter extends the RDD and BJD filter to the time-varying case.

Remark 4
In the limit, by using Lemma 4.2 and that $\operatorname{tr}(\hat{H}_i C P_i C^T \hat{H}_i)$ is finite [12], the robust multiple-fault detection filter problem satisfies

$$\frac{E[\hat{h}_i(t)\hat{h}_i(t)^T]}{E[h_i(t)h_i(t)^T]} \to 0$$

for $i = 1 \cdots q$. This implies that the transmissions from the associated nuisance faults to their associated projected residuals are zero.

4.3. *Condition on fault detection and identification*

In this section, three conditions to ensure that the faults can be detected and identified are assumed. First, $C\mathscr{T}_1 \cdots C\mathscr{T}_q$ are independent. If they are not independent, different faults will produce the same non-zero projected residuals and therefore the faults cannot be identified. This

is equivalent to the output separability condition in Reference [3]. Note that $C\mathscr{T}_1 \cdots C\mathscr{T}_q$ are independent if and only if $C\mathscr{W}_1 \cdots C\mathscr{W}_q$ are independent.

The other two conditions are assumed for time-invariant systems only. The first condition is that the invariant zeros of $(C, A, [F_1 \cdots F_q])$ are either the invariant zeros of (C, A, F_i), $i = 1 \cdots q$, or in the left-half plane. This is from the mutually detectable condition for the RDD filter because the robust multiple-fault detection filter becomes equivalent to the RDD filter in the limit. $F_1 \cdots F_q$ are mutually detectable if $(C, A, [F_1 \cdots F_q])$ does not have more invariant zeros than (C, A, F_i), $i = 1 \cdots q$ [3]. If $F_1 \cdots F_q$ are not mutually detectable, the extra invariant zeros will become part of the eigenvalues of the detection filter. If the extra invariant zeros are in the right-half plane, no stable detection filter can be found to isolate these q faults. A numerical example is given in Section 6.2.3. The second condition is that (C, A, F_i) cannot have invariant zeros at the origin if μ_i needs to be detected [12]. This ensures a non-zero projected residual in steady state when its associated target fault occurs.

5. MINIMIZATION WITH RESPECT TO $\hat{H}_1 \cdots \hat{H}_s$

In this section, the robust multiple-fault detection filter problem is solved with $\hat{H}_1 \cdots \hat{H}_s$ derived from solving the minimization problem instead of defined *a priori* by (9). From (11), by using $\hat{H}_i = \rho_i \rho_i^T$, the minimization problem becomes

$$\min_{L, \hat{H}_1 \cdots \hat{H}_s} \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} \operatorname{tr}\left[\sum_{i=1}^{s} \rho_i^T C(W_i + P_i) C^T \rho_i \right] \mathrm{d}t$$

subject to (12) and $\rho_i^T \rho_i = I_{m_i}$ where m_i is the rank of (9). By using matrix Lagrange multipliers K_i and Σ_i to form the variational Hamiltonian, the first-order necessary conditions imply that the optimal solution for L and the dynamics of K_i are still (14) and (15), respectively. Further, from the first-order necessary condition $C(W_i + P_i)C^T \rho_i = \rho_i \Sigma_i$, the optimal solution for $\hat{H}_i$ is

$$\hat{H}_i^* = [\rho_{i,1} \quad \rho_{i,2} \quad \cdots \quad \rho_{i,m_i}][\rho_{i,1} \quad \rho_{i,2} \quad \cdots \quad \rho_{i,m_i}]^{\mathrm{T}} \tag{26}$$

where $\rho_{i,1} \cdots \rho_{i,m_i}$ are the eigenvectors of $C(W_i + P_i)C^T$ associated with the smallest m_i eigenvalues. To obtain the optimal solutions for L and $\hat{H}_1 \cdots \hat{H}_s$, (12), (14), (15) and (26) have to be solved simultaneously. For the infinite-time case, (14), (17), (19) and (26) have to be solved simultaneously. In Section 4.2, it is shown that (9) minimizes the cost criterion in the limit. Therefore, (26) becomes equivalent to (9) in the limit. Note that, for time-invariant systems, (9) is the projector used by the RDD filter [3].

6. EXAMPLE

In this section, two numerical examples are used to demonstrate the robust multiple-fault detection filter. In Section 6.1, the filters are derived in the forms of unknown input observer, BJD filter and RDD filter, respectively. In Section 6.2, the filters are derived to show that the filter has behaviours similar to the RDD and BJD filters.

6.1. Example 1

In this section, a linear time-invariant system for the F16XL aircraft [6] is used to demonstrate the performance of the robust multiple-fault detection filter. The system has four states (longitudinal velocity x_u, normal velocity x_w, pitch rate x_q and pitch angle x_θ), one control input (elevon deflection angle u_δ), four measurements (longitudinal velocity y_u, normal velocity y_w, pitch rate y_q and pitch angle y_θ) and one disturbance input (wind gust u_{wg}). The system matrices are

$$A = \begin{bmatrix} -0.0674 & 0.0430 & -0.8886 & -0.5587 \\ 0.0205 & -1.4666 & 16.5800 & -0.0299 \\ 0.1377 & -1.6788 & -0.6819 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad B_\delta = \begin{bmatrix} -0.1672 \\ -1.5179 \\ -9.7842 \\ 0 \end{bmatrix},$$

$$B_{\text{wg}} = \begin{bmatrix} 0.0430 \\ -1.4666 \\ -1.6788 \\ 0 \end{bmatrix}, \quad C = I$$

Three faults in pitch angle sensor y_θ, elevon deflector u_δ and wind gust u_{wg} are considered. In this example, the wind gust is considered as a fault instead of a process noise. The fault directions are [4]

$$F_\theta = \begin{bmatrix} 0 & -0.5587 \\ 0 & -0.0299 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad F_\delta = \begin{bmatrix} -0.1672 \\ -1.5179 \\ -9.7842 \\ 0 \end{bmatrix}, \quad F_{\text{wg}} = \begin{bmatrix} 0.0430 \\ -1.4666 \\ -1.6788 \\ 0 \end{bmatrix}$$

In Section 6.1.1, the filters are derived in the form of unknown input observer where $s = 1$. In Section 6.1.2, the filter is derived in the form of the BJD filter where $s = 3$. In Section 6.1.3, the filter is derived in the form of the RDD filter where $s = 2$. In Section 6.1.4, the filter is derived to show that the sensitivity of the projected residuals to their associated target faults can be enhanced.

6.1.1. Unknown input observer

In this section, the filters are derived in the form of unknown input observer where $s = 1$. Since each filter can detect only one fault, three filters are needed. Let $F_1 = F_\theta$, $F_2 = F_\delta$ and $F_3 = F_{\text{wg}}$. The weightings are chosen as $\gamma_1 = \gamma_2 = \gamma_3 = 10^{-6}$, $Q_1 = 0.1I$, $Q_2 = Q_3 = 1$ and $V = I$. The steady-state solutions of (13) are obtained for $i = 1 \cdots 3$, respectively. Then, three single-fault filters (3) are obtained by (20). Figure 1 shows the frequency response from each fault to the projected residual $\hat{H}_i r$ (4) of each filter. Note that each filter has only one projected residual $\hat{H}_i r$ for detecting the fault F_i. The projectors $\hat{H}_1 \cdots \hat{H}_3$ are defined by (9). The dashed line represents the pitch angle sensor fault. The dashdot line represents the elevon deflector fault. The solid line

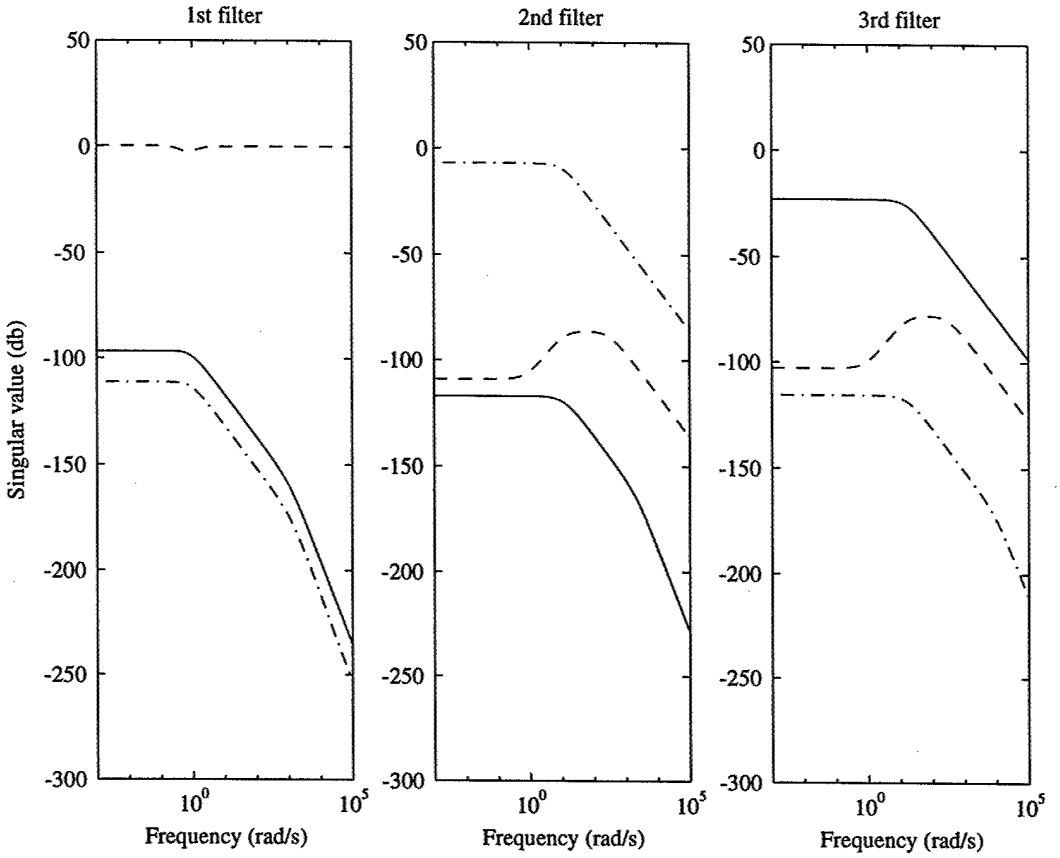


Figure 1. Frequency response of the three single-fault filters when $s = 1$.

represents the wind gust fault. This example shows that the projected residual of each filter is only sensitive to its associated target fault, but not to its associated nuisance fault.

6.1.2. Beard–Jones detection filter

In this section, the filter is derived in the form of the BJD filter where $s = 3$. Since the filter can detect all three faults, only one filter is needed. The filter gain, satisfying (17), (18) and (19), is obtained by using the gradient method to solve (16) numerically with $\hat{H}_1 \cdots \hat{H}_3$ defined *a priori* by (9). Figure 2 shows the frequency response from each fault to the three projected residuals $\hat{H}_1 r \cdots \hat{H}_3 r$ (4) of the filter (3). This example shows that one multiple-fault filter works as well as three single-fault filters.

6.1.3. Restricted diagonal detection filter

Since the wind gust is a disturbance, it does not need to be detected, but only needs to be blocked. Therefore, in this section, the filter is derived in the form of the RDD filter where $s = 2$. The filter gain, satisfying (17), (18) and (19), is obtained by using the gradient method to solve

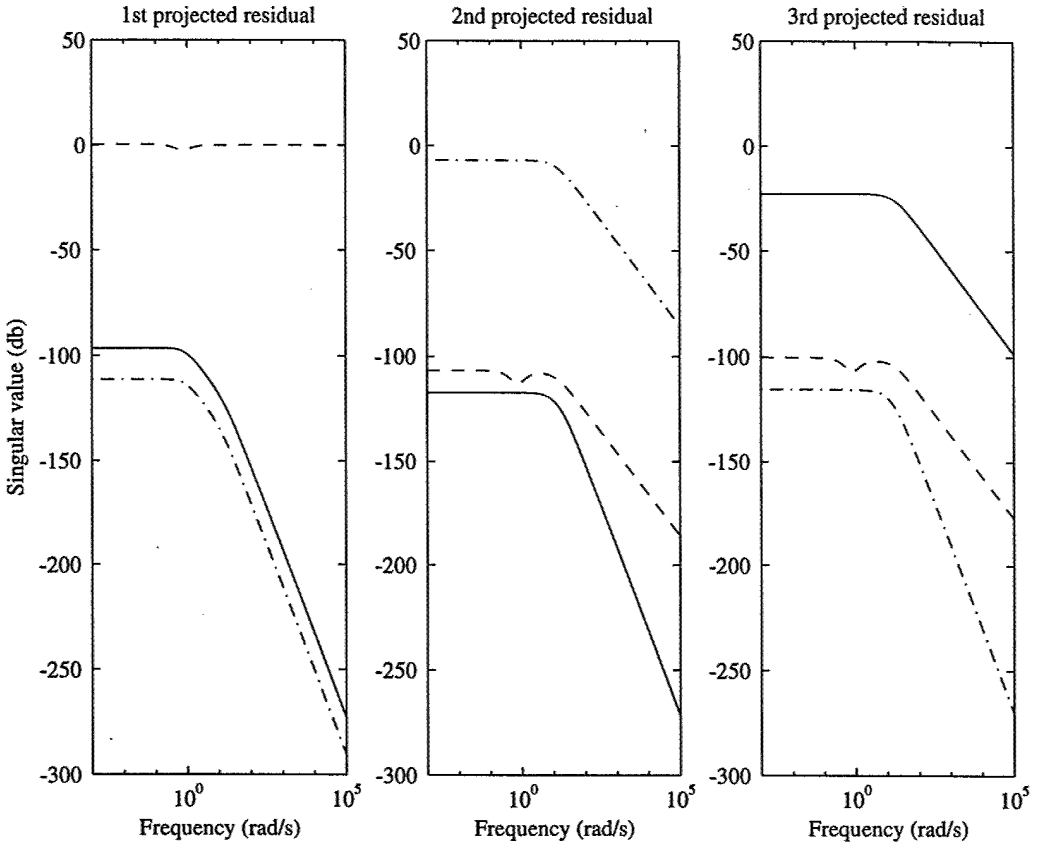


Figure 2. Frequency response of the multiple-fault filter when $s = 3$.

(16) numerically with $\hat{H}_1$ and $\hat{H}_2$ defined *a priori* by (9). In Figure 3, the left and middle figures show the frequency response from each fault to the two projected residuals, $\hat{H}_1 r$ and $\hat{H}_2 r$ (4), of the filter (3). Note that the filter has only two projected residuals because only two faults, F_1 and F_2, are detected. These two figures show that the pitch angle sensor fault and elevon deflector fault can still be detected and identified even though $s = 2$. To compare with the filter derived in the previous example where $s = 3$, the right figure in Figure 3 shows the frequency response from each fault to the projected residual $\hat{H}_3 r$ used for detecting the wind gust fault in previous example. This figure shows that the wind gust fault can no longer be identified from the other two faults. This example shows that the multiple-fault filter still works well after relaxing the constraint on detecting the wind gust fault.

6.1.4. Enhancement of associated target fault sensitivity

In this section, another filter in the form of the RDD filter where $s = 2$ is derived to show that the sensitivity of the projected residuals to their associated target faults can be enhanced. The weightings are the same except $Q_1 = 0.64I$ and $Q_2 = 4.73$. In Figure 4, the performance of this

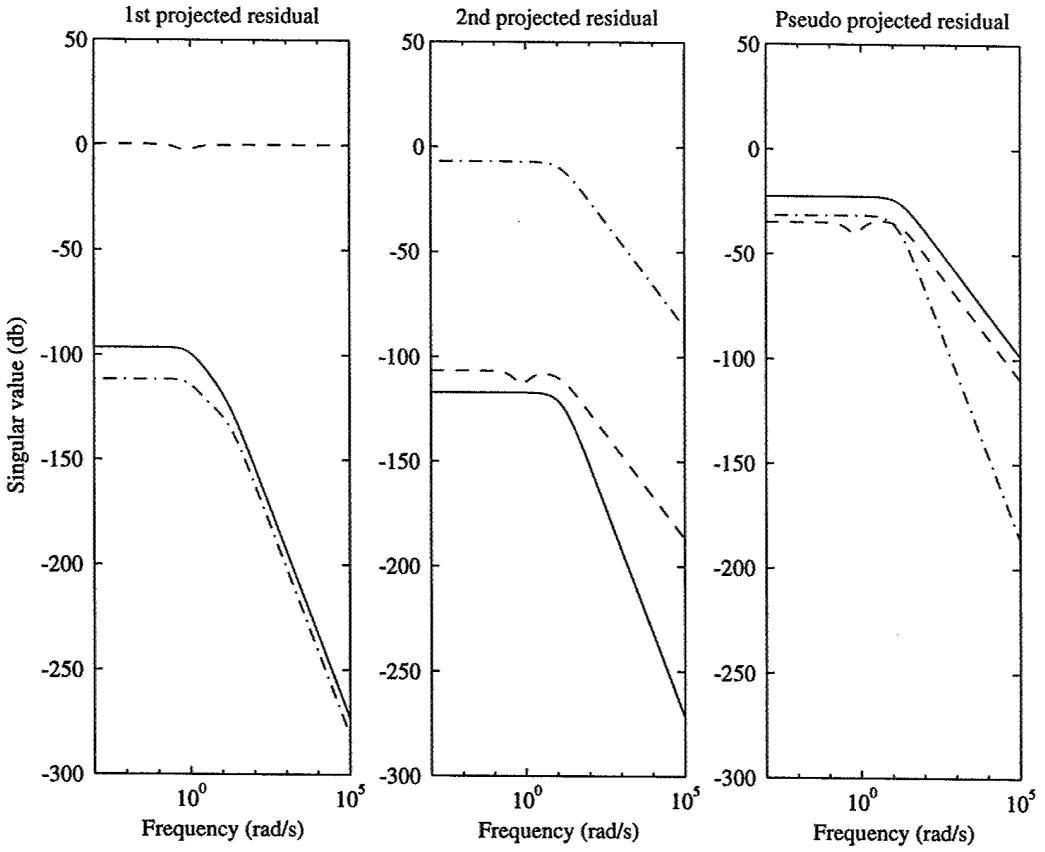


Figure 3. Frequency response of the multiple-fault filter when $s = 2$.

filter is compared to the filter derived in the previous example. The left figure shows the frequency response from the pitch angle sensor fault to its associated projected residuals when $Q_1 = 0.1I$ and $0.64I$, respectively. The right figure shows the frequency response from the elevon deflector fault to its associated projected residuals when $Q_2 = 1$ and 4.73, respectively. This example shows that the sensitivity of the projected residuals to their associated target faults can be enhanced by increasing the weightings of the associated target faults.

6.2. Example 2

In this section, three numerical examples are used to show that the robust multiple-fault detection filter has behaviours similar to the RDD and BJD filters. In Section 6.2.1, the filter is derived when the fault has an invariant zero in the right-half plane. In Section 6.2.2, the filter is derived when the fault has an invariant zero in the left-half plane. In Section 6.2.3, the filter is derived when the faults are not mutually detectable.

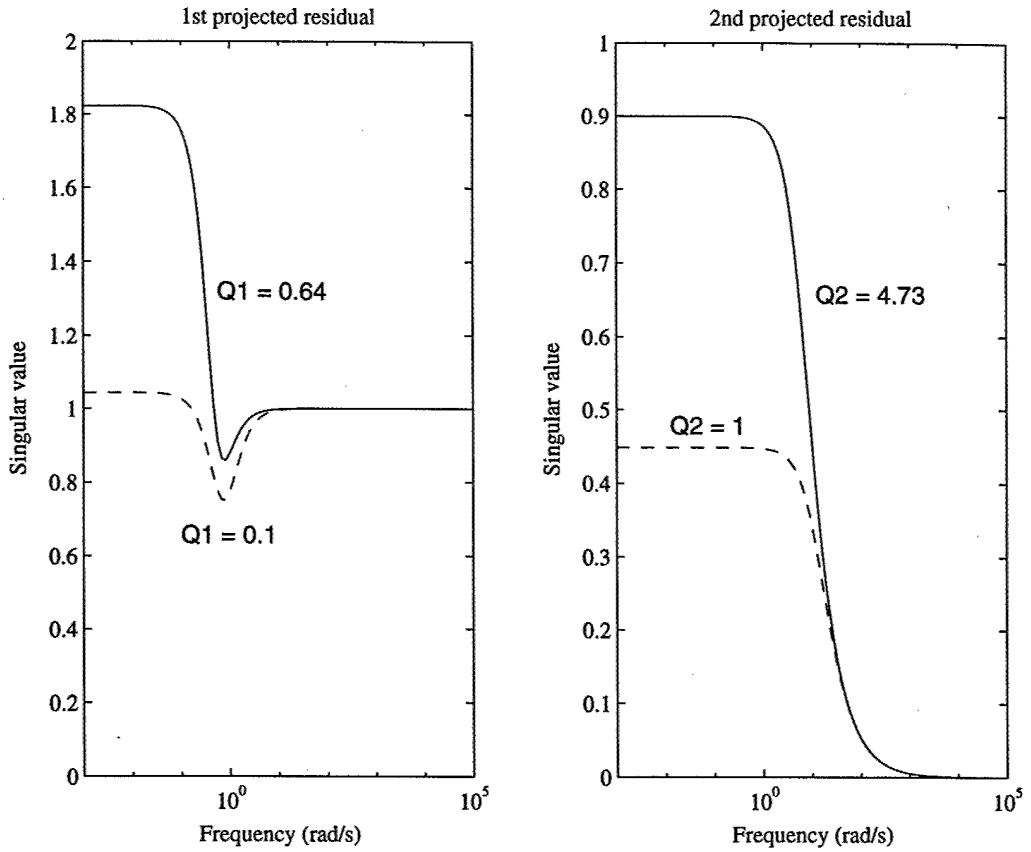


Figure 4. Frequency response of the multiple-fault filter when $s = 2$.

6.2.1. Right-half-plane invariant zero.
Consider the time-invariant system from Reference [4],

$$A = \begin{bmatrix} 0 & 3 & 4 \\ 1 & 2 & 3 \\ 0 & 2 & 5 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad F_1 = \begin{bmatrix} 1 \\ -0.5 \\ 0.5 \end{bmatrix}, \quad F_2 = \begin{bmatrix} -3 \\ 1 \\ 0 \end{bmatrix}$$

There is no process noise. (C, A, F_2) has an invariant zero at 3 and the invariant zero direction is $v = [1\ 0\ 0]^T$. By using (21), $\mathscr{T}_1 = \mathrm{Im}\, F_1$ and $\mathscr{T}_2 = \mathrm{Im}[F_2\ v]$. Since $\mathscr{T}_1 \oplus \mathscr{T}_2 = \mathscr{X}$, there is no complementary subspace.

A multiple-fault filter is derived similarly as before to detect and identify these two faults. The weightings are chosen as $\gamma_1 = \gamma_2 = 10^{-6}$, $Q_1 = Q_2 = 0.25$ and $V = I$. The eigenvectors of the filter are very close to $\mathscr{T}_1$ and $\mathscr{T}_2$ similar to the BJD filter. Since the invariant zero direction is approximately included in the invariant subspace of F_2 generated by the filter, none of the

eigenvalues of the filter is close to the invariant zero at 3 [3]. The eigenvalues of the filter are -0.5865, -5.3789 and -7.1102.

6.2.2. Left-half-plane invariant zero

Consider the same time-invariant system from Section 6.2.1 except $F_2 = [3\ 1\ 0]^T$. (C, A, F_2) has an invariant zero at -3 and the invariant zero direction is $v = [1\ 0\ 0]^T$. By using (21), $\mathscr{T}_1 = \mathrm{Im}\, F_1$ and $\mathscr{T}_2 = \mathrm{Im}[F_2 v]$. Since $\mathscr{T}_1 \oplus \mathscr{T}_2 = \mathscr{X}$, there is no complementary subspace.

A multiple-fault filter is derived with the same weightings as in Section 6.2.1. The eigenvectors of the filter are very close to $\mathscr{T}_1$ and $\mathscr{T}_2$ similar to the BJD filter. Since the invariant zero direction is approximately included in the invariant subspace of F_2 generated by the filter, none of the eigenvalues of the filter is close to the invariant zero at -3 [3]. The eigenvalues of the filter are -0.5865, -5.3789 and -7.1102.

Remark 5

For the single-fault filter [12], the invariant zero directions associated with the left-half-plane invariant zeros are not included in the invariant subspace and part of the eigenvalues of the filter are very close to the invariant zeros. Although the invariant zero directions associated with the right-half-plane invariant zeros are included in the invariant subspace, part of the eigenvalues of the filter are very close to the mirror images of the invariant zeros. To avoid this situation, the fault directions have to be modified. However, as demonstrated by the numerical examples in Sections 6.2.1 and 6.2.2, the multiple-fault filter automatically includes the invariant zero directions in the invariant subspaces and none of the eigenvalues of the filter is close to the invariant zeros or their mirror images.

6.2.3. Non-mutually detectable faults

Consider the same time-invariant system from Section 6.2.1 except $F_2 = [5\ 1\ 1]^T$. F_1 and F_2 are not mutually detectable because $(C, A, [F_1\ F_2])$ has an invariant zero at -1.5 while (C, A, F_1) and (C, A, F_2) do not have any invariant zero. By using (21), $\mathscr{T}_1 = \mathrm{Im}\, F_1$ and $\mathscr{T}_2 = \mathrm{Im}\, F_2$. Since $\mathscr{T}_1 \oplus \mathscr{T}_2 \subset \mathscr{X}$, there is a complementary subspace.

A multiple-fault filter is derived with the same weightings as in Section 6.2.1. Two of the eigenvectors of the filter are very close to $\mathscr{T}_1$ and $\mathscr{T}_2$ similar to the BJD filter. Since F_1 and F_2 are not mutually detectable, one of the eigenvalues of the filter is very close to the extra invariant zero at -1.5 [3]. The eigenvalues of the filter are -1.5008, -5.7648 and -6.8185.

Remark 6

A multiple-fault filter is also derived for two non-mutually detectable faults where the extra invariant zero is in the right-half plane. Although a stable filter can be derived numerically by minimizing the cost criterion, the minimal cost is large and the filter cannot isolate the faults. This is consistent with the BJD filter in that the extra invariant zero will become one of the eigenvalues of the filter if the filter generates the invariant subspaces to isolate the faults [3]. Therefore, it is impossible to obtain a stable multiple-fault filter that can isolate the faults. However, two single-fault filters can be used to monitor these two faults.

7. CONCLUSION

Different from other design algorithms for the RDD or BJD filter which explicitly force the geometric structure by using eigenstructure assignment or geometric theory, the robust multiple-fault detection filter is derived from solving a stochastic minimization problem and only in the limit, is the geometric structure of the RDD filter recovered and the faults are completely isolated. When it is not in the limit, the filter only isolates the faults within approximate unobservable subspaces. This new feature allows the filter to be potentially more robust because of the additional design freedom which allows different degrees of fault isolation. Furthermore, a mechanism that enhances the sensitivity of the projected residuals to their associated target faults is provided. Finally, the filter can be applied to time-varying systems. Although the process of deriving the filter gain requires the solution to a two-point boundary value problem, the filter gain computation can be done off-line so that the filter implementation is as straightforward as the RDD filter. However, further research is needed in developing a numerical algorithm to solve the optimization problem more efficiently.

APPENDIX A

A.1. Proof of Lemma 4.1

To show (24a), for $i = 1$, by using Lemma A.1 in Appendix A.5

$$\Gamma^{-1}\left(\sum_{j=1}^{q} H_j\right)^{-1} H_1\Gamma = \left(\sum_{j=1}^{q} \Gamma^T H_j \Gamma\right)^{-1} \Gamma^T H_1 \Gamma$$

$$= \begin{bmatrix} \bar{H}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \bar{H}_q \end{bmatrix}^{-1} \begin{bmatrix} \bar{H}_1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

$$\Gamma^{-1}\left(\sum_{j=1}^{q} K_j\right)^{-1} K_1\Gamma = \left(\sum_{j=1}^{q} \Gamma^T K_j \Gamma\right)^{-1} \Gamma^T K_1 \Gamma$$

$$= \begin{bmatrix} \bar{K}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \bar{K}_q \end{bmatrix}^{-1} \begin{bmatrix} \bar{K}_1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

Therefore, $(\sum_{j=1}^{q} H_j)^{-1} H_1 = (\sum_{j=1}^{q} K_j)^{-1} K_1$. It can be shown similarly for the cases where $i = 2 \cdots q$. This completes the proof for (24a).

To show (24b), by substituting (24a) into (14)

$$L^* = \left(\sum_{i=1}^{s} H_i\right)^{-1} \left[\sum_{i=1}^{s} H_i (P_i + W_i)\right] C^T V^{-1} \tag{A1}$$

By multiplying (12) by H_i from the left and right, substituting (A1) and using Lemma A.2 in Appendix A.5

$$H_i[-\dot{W}_i + (A - P_iC^TV^{-1}C)W_i + W_i(A - P_iC^TV^{-1}C)^T - W_iC^TV^{-1}CW_i]H_i = 0$$

Note that, from (20), $A - P_iC^TV^{-1}C$ is the closed-loop A matrix of the filter when only the fault F_i is detected. Then

$$\dot{W}_i = (A - P_iC^TV^{-1}C)W_i + W_i(A - P_iC^TV^{-1}C)^T - W_iC^TV^{-1}CW_i + \hat{T}_i\hat{T}_i^T$$

where $\operatorname{Im}\hat{T}_i = \hat{\mathscr{T}}_i$ because $\operatorname{Ker}H_i = \hat{\mathscr{T}}_i$. Since $\hat{\mathscr{T}}_i$ is $(A - P_iC^TV^{-1}C)$-invariant [12], the controllable subspace of $(A - P_iC^TV^{-1}C, \hat{T}_i)$ is $\hat{\mathscr{T}}_i$ and $\operatorname{Im}W_i = \hat{\mathscr{T}}_i$. Since $\operatorname{Ker}H_i = \hat{\mathscr{T}}_i$, $H_iW_i = 0$. This completes the proof for (24b).

A.2. Proof of Lemma 4.2

By multiplying (12) by H_i from the left and right, substituting (25) and using Lemma A.2 in Appendix A.5

$$H_i[-\dot{W}_i + (A - P_iC^TV^{-1}C)W_i + W_i(A - P_iC^TV^{-1}C)^T]H_i = 0$$

Then

$$\dot{W}_i = (A - P_iC^TV^{-1}C)W_i + W_i(A - P_iC^TV^{-1}C)^T + \hat{T}_i\hat{T}_i^T$$

where $\operatorname{Im}\hat{T}_i = \hat{\mathscr{T}}_i$ because $\operatorname{Ker}H_i = \hat{\mathscr{T}}_i$. Since $\hat{\mathscr{T}}_i$ is $(A - P_iC^TV^{-1}C)$-invariant [12], the controllable subspace of $(A - P_iC^TV^{-1}C, \hat{T}_i)$ is $\hat{\mathscr{T}}_i$. Then, the image of the controllability grammian W_i is $\hat{\mathscr{T}}_i$. Since $\operatorname{Ker}\hat{H}_i = C\hat{\mathscr{T}}_i$ from (9), $\hat{H}_iCW_iC^T\hat{H}_i = 0$. Therefore

$$J^* = \frac{1}{t_1 - t_0}\int_{t_0}^{t_1} \operatorname{tr}\left(\sum_{i=1}^{s}\hat{H}_iCW_iC^T\hat{H}_i\right)\mathrm{d}t = 0$$

A.3. Proof of Theorem 4.3

Since P_i goes to infinity in the limit, $\Pi_i \triangleq P_i^{-1}$ has a null space [12] and

$$-\dot{\Pi}_i = \Pi_iA + A^T\Pi_i + \Pi_i\left(\frac{1}{\gamma_i}\hat{F}_i\hat{Q}_i\hat{F}_i^T - F_iQ_iF_i^T + B_wQ_wB_w^T\right)\Pi_i - C^TV^{-1}C \tag{A2}$$

When the associated nuisance fault occurs, the dynamic equation of the error without process and sensor noises can be written as

$$\Pi_i\dot{e} = \Pi_i(A - LC)e + \Pi_i\hat{F}_i\hat{\mu}_i$$

By adding $\dot{\Pi}_ie$ to both sides and substituting (A2)

$$\frac{\mathrm{d}}{\mathrm{d}t}(\Pi_ie) = -\left[\Pi_iLC + A^T\Pi_i + \Pi_i\left(\frac{1}{\gamma_i}\hat{F}_i\hat{Q}_i\hat{F}_i^T - F_iQ_iF_i^T + B_wQ_wB_w^T\right)\right.$$
$$\left.\times\,\Pi_i - C^TV^{-1}C\right]e + \Pi_i\hat{F}_i\hat{\mu}_i \tag{A3}$$

Let $\bar{\Pi}_i = \lim_{\gamma\to 0}\Pi$. Since $\operatorname{Ker}\bar{\Pi}_i = \hat{\mathscr{T}}_i$ [12]

$$\bar{\Pi}_i\left(\sum_{k=1}^{q}H_k\right)^{-1}H_j = \begin{cases}\bar{\Pi}_i, & i = j\\ 0, & i \neq j\end{cases} \tag{A4}$$

which can be shown similarly to Lemma A.2 in Appendix A.5. In the limit, by substituting (25) and (A4) into (A3)

$$\frac{\mathrm{d}}{\mathrm{d}t}(\bar{\Pi}_i e) = -\left[A^T + \bar{\Pi}_i\left(\frac{1}{\gamma_i}\hat{F}_i\hat{Q}_i\hat{F}_i^T - F_iQ_iF_i^T + B_wQ_wB_w^T\right)\right]\bar{\Pi}_i e + \bar{\Pi}_i\hat{F}_i\hat{\mu}_i \tag{A5}$$

If the error initially lies in $\mathrm{Ker}\,\bar{\Pi}_i$, (A5) implies that the error will never leave $\mathrm{Ker}\,\bar{\Pi}_i$ because $\bar{\Pi}_i\hat{F}_i = 0$ [12]. Therefore, $\mathrm{Ker}\,\bar{\Pi}_i$ is $(A - LC)$-invariant where L is in (25). Since $\mathrm{Ker}\,\bar{\Pi}_i = \hat{\mathscr{T}}_i$ [12], $\hat{\mathscr{T}}_i$ is $(A - LC)$-invariant where L is in (25).

A.4. Proof of Corollary 4.4

When $s = q$, $\mathscr{T}_i = \hat{\mathscr{T}}_1 \cap \cdots \cap \hat{\mathscr{T}}_{i-1} \cap \hat{\mathscr{T}}_{i+1} \cap \cdots \cap \hat{\mathscr{T}}_q$. From Theorem 4.3, $\hat{\mathscr{T}}_1 \cdots \hat{\mathscr{T}}_q$ are $(A - LC)$-invariant where L is in (25). Therefore, $\mathscr{T}_1 \cdots \mathscr{T}_q$ are $(A - LC)$-invariant where L is in (25).

A.5. Lemmas

Lemma A.1
There exists a state transformation Γ

$$[\mathscr{T}_1 \ \cdots \ \mathscr{T}_q] = \Gamma\begin{bmatrix} Z_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & Z_q \end{bmatrix}$$

where Z_i, $i = 1 \cdots q$, are any invertible matrices with dimension equivalent to $\dim \mathscr{T}_i$ such that $K_1 \cdots K_q$ are in the form of

$$\Gamma^T K_1 \Gamma = \begin{bmatrix} \bar{K}_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \Gamma^T K_2 \Gamma = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \bar{K}_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \cdots \quad \Gamma^T K_q \Gamma = \begin{bmatrix} 0 & 0 \\ 0 & \bar{K}_q \end{bmatrix}$$

in the limit where $\bar{K}_1 \cdots \bar{K}_q$ are invertible and $H_1 \cdots H_q$ are in the form of

$$\Gamma^T H_1 \Gamma = \begin{bmatrix} \bar{H}_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \Gamma^T H_2 \Gamma = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \bar{H}_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \cdots \quad \Gamma^T H_q \Gamma = \begin{bmatrix} 0 & 0 \\ 0 & \bar{H}_q \end{bmatrix}$$

where $\bar{H}_1 \cdots \bar{H}_q$ are invertible.

Proof
Since $\mathrm{Ker}\,\hat{H}_i = C\hat{\mathscr{T}}_i$ from (9) and $\hat{\mathscr{T}}_i$ is $(A - LC)$-invariant in the limit by the assumption, the unobservable subspace of $(\hat{H}_iC, A - LC)$ is $\hat{\mathscr{T}}_i$. Then, from the Lyapunov equation (15), the null space of the observability grammian K_i is $\hat{\mathscr{T}}_i$ in the limit. For $i = 1$

$$\mathrm{Ker}\,K_1 = \hat{\mathscr{T}}_1 = \Gamma\begin{bmatrix} 0 \\ \hat{Z}_1 \end{bmatrix} \Rightarrow K_1\Gamma\begin{bmatrix} 0 \\ \hat{Z}_1 \end{bmatrix} = 0 \Rightarrow \Gamma^T K_1 \Gamma\begin{bmatrix} 0 \\ \hat{Z}_1 \end{bmatrix} = 0$$

where $\hat{Z}_1 = \text{diag}(Z_2 \cdots Z_q)$. Since $\hat{Z}_1$ is invertible and $\Gamma^T K_1 \Gamma$ is symmetric

$$\Gamma^T K_1 \Gamma = \begin{bmatrix} \bar{K}_1 & 0 \\ 0 & 0 \end{bmatrix}$$

It can be shown similarly for $K_2 \cdots K_q$ and $H_1 \cdots H_q$. □

Lemma A.1

$$H_i \left(\sum_{k=1}^{q} H_k \right)^{-1} H_j = \begin{cases} H_i, & i = j \\ 0, & i \neq j \end{cases}$$

Proof
For $i = j = 1$, by using Lemma A.2,

$$\Gamma^T H_1 \left(\sum_{k=1}^{q} H_k \right)^{-1} H_1 \Gamma = (\Gamma^T H_1 \Gamma) \left(\sum_{k=1}^{q} \Gamma^T H_k \Gamma \right)^{-1} (\Gamma^T H_1 \Gamma)$$

$$= \begin{bmatrix} \bar{H}_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{H}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \bar{H}_q \end{bmatrix}^{-1} \begin{bmatrix} \bar{H}_1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{H}_1 & 0 \\ 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \bar{H}_1 & 0 \\ 0 & 0 \end{bmatrix} = \Gamma^T H_1 \Gamma$$

Therefore, $H_1 (\sum_{k=1}^{q} H_k)^{-1} H_1 = H_1$. It can be shown similarly for other cases where $i = j$. For $i = 1$ and $j = 2$

$$\Gamma^T H_1 \left(\sum_{k=1}^{q} H_k \right)^{-1} H_2 \Gamma = (\Gamma^T H_1 \Gamma) \left(\sum_{k=1}^{q} \Gamma^T H_k \Gamma \right)^{-1} (\Gamma^T H_2 \Gamma)$$

$$= \begin{bmatrix} \bar{H}_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{H}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \bar{H}_q \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 & 0 \\ 0 & \bar{H}_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & \bar{H}_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} = 0$$

Therefore, $H_1 (\sum_{k=1}^{q} H_k)^{-1} H_2 = 0$. It can be shown similarly for other cases where $i \neq j$. □

REFERENCES

1. Beard RV. Failure accommodation in linear systems through self-reorganization. *Ph.D. Thesis*, Massachusetts Institute of Technology, 1971.
2. Jones HL. Failure detection in linear systems. *Ph.D. Thesis*, Massachusetts Institute of Technology, 1973.
3. Massoumnia M-A. A geometric approach to the synthesis of failure detection filters. *IEEE Transactions on Automatic Control* 1986; **AC-31**(9):839–846.
4. White JE, Speyer JL. Detection filter design: spectral theory and algorithms. *IEEE Transactions on Automatic Control* 1987; **AC-32**(7):593–603.
5. Douglas RK, Speyer JL. $\mathscr{H}_\infty$ bounded fault detection filter. *AIAA Journal of Guidance, Control, and Dynamics* 1999; **22**(1):129–138.
6. Douglas RK, Speyer JL. Robust fault detection filter design. *AIAA Journal of Guidance, Control, and Dynamics* 1996; **19**(1):214–218.
7. Massoumnia MA, Verghese GC, Willsky AS. Failure detection and identification. *IEEE Transactions on Automatic Control* 1989; **AC-34**(3):316–321.
8. Frank PM. Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy—a survey and some new results. *Automatica* 1990; **26**(3):459–474.
9. Patton RJ, Chen J. Robust fault detection of jet engine sensor systems using eigenstructure assignment. *AIAA Journal of Guidance, Control, and Dynamics* 1992; **15**(6):1491–1497.
10. Chung WH, Speyer JL. A game theoretic fault detection filter. *IEEE Transactions on Automatic Control* 1998; **AC-43**(2):143–161.
11. Chen RH, Speyer JL. A generalized least-squares fault detection filter. *International Journal of Adaptive Control and Signal Processing—Special Issue: Fault Detection and Isolation* 2000; **14**(7):747–757.
12. Chen RH, Lewis Mingori D, Speyer JL. Optimal stochastic fault detection filter. *Automatica*, to be published.
13. Douglas RK, Chen RH, Speyer JL. Model input reduction. In *Proceedings of the American Control Conference* 1997; 3882–3886.
14. Bell DJ, Jacobson DH. *Singular Optimal Control Problems*. Academic Press: New York, 1975.
15. Moylan PJ, Moore JB. Generalizations of singular optimal control theory. *Automatica* 1971; **7**(5):591–598.
16. Athans M. The matrix minimum principle. *Information and Control*, 1968; **11**(5/6):592–606.
17. Chen RH. Fault detection filters for robust analytical redundancy. *Ph.D. Thesis*, University of California at Los Angeles, 2000.

Appendix G

"Optimal Stochastic Fault Detection Filter,"

Robert H. Chen, D. Lewis Mingori and Jason L. Speyer,

***Automatica,* vol. 39 (2003) 377-390.**

Available online at www.sciencedirect.com
SCIENCE DIRECT
PERGAMON
Automatica 39 (2003) 377–390
automatica
www.elsevier.com/locate/automatica

Optimal stochastic fault detection filter☆

Robert H. Chen[a], D. Lewis Mingori[a], Jason L. Speyer[a,*]

[a] *Mechanical and Aerospace Engineering Department, University of California, Los Angeles, CA 90095, USA*

Received 12 May 1999; received in revised form 17 March 2002; accepted 11 September 2002

Abstract

A fault detection and identification algorithm, called optimal stochastic fault detection filter, is determined. The objective of the filter is to detect a single fault, called the target fault, and block other faults, called the nuisance faults, in the presence of the process and sensor noises. The filter is derived by maximizing the transmission from the target fault to the projected output error while minimizing the transmission from the nuisance faults. Therefore, the residual is affected primarily by the target fault and minimally by the nuisance faults. The transmission from the process and sensor noises is also minimized so that the filter is robust with respect to these disturbances. It is shown that the filter recovers the geometric structure of the unknown input observer in the limit where the weighting on the nuisance fault transmission goes to infinity. Further, the asymptotic behavior of the filter near the limit is determined by using a perturbation method. Filter designs can be obtained for both time-invariant and time-varying systems.

Keywords: Fault detection and identification; Analytical redundancy; Unknown input observer; Robust fault detection filter; Time-varying system; Perturbation theory

1. Introduction

Any system under automatic control demands a high degree of system reliability. This requires a health monitoring system capable of detecting any plant, actuator and sensor faults as they occur and identifying the faulty components. One approach, analytical redundancy which reduces the need for hardware redundancy, uses the modeled dynamic relationship between system inputs and measured system outputs to form a residual process which can be used for detecting and identifying faults. A popular approach to analytical redundancy is the unknown input observer (Chen & Speyer, 2000; Chung & Speyer, 1998; Frank, 1990; Massoumnia, Verghese, & Willsky, 1989; Patton & Chen, 1992) which divides the faults into two groups: a single target fault and possibly several nuisance faults. The nuisance faults are placed in an invariant subspace which is unobservable to the residual. Therefore, the residual is only sensitive to the target fault, but not to the nuisance faults.

In this paper, a design algorithm, called optimal stochastic fault detection filter, is determined for the unknown input observer. The filter is derived by maximizing the transmission from the target fault while minimizing the transmission from the nuisance faults. The transmission is defined on the projected output error by using a projector to be derived from solving the optimization problem. Therefore, the residual is affected primarily by the target fault and minimally by the nuisance faults. The transmission from the process and sensor noises is also minimized so that the filter is robust with respect to these disturbances. Since certain types of model uncertainties can be modeled as additive noises (Patton & Chen, 1992; Douglas, Chen & Speyer, 1997) the filter can also be made robust to these model uncertainties.

In the limit where the weighting on the nuisance fault transmission goes to infinity, the filter blocks the nuisance faults completely. It is shown that the filter places the nuisance faults into a minimal (C,A)-unobservability subspace for time-invariant systems and a similar invariant subspace for time-varying systems. Therefore, the filter recovers the geometric structure of the unknown input observer in the limit and extends the unknown input observer to the

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Rene Boel under the direction of Editor Tamer Basar.

* Corresponding author. Tel.: +1-310-206-4451; fax: +1-310-206-2302.

E-mail addresses: chrobert@talus.seas.ucla.edu (R.H. Chen), mingori@seas.ucla.edu (D. L. Mingori), speyer@seas.ucla.edu (J.L. Speyer).

0005-1098/03/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0005-1098(02)00245-5

time-varying case similar to Chen and Speyer (2000) and Chung and Speyer (1998). These limiting results are important in ensuring that both fault detection and identification can occur. For time-invariant systems, the nuisance fault directions are generalized to prevent the invariant zeros of the nuisance faults or their mirror images from becoming part of the eigenvalues of the filter.

The behavior of the filter near and in the limit can be determined by using a perturbation method. In particular, the perturbation method captures the asymptotic behavior of the Riccati equation that defines the filter gain and generalizes the result of Kwakernaak and Sivan (1972). Note that Chen and Speyer (2000) and Chung and Speyer (1998) use the Goh transformation in singular optimal control theory (Bell & Jacobson, 1975; Moylan & Moore, 1971) to determine the filter in the limit. Although the Goh transformation cannot determine the asymptotic behavior of the filter near the limit, it is shown that it produces a limiting Riccati equation which is the same as that determined from the perturbation method. Finally, the asymptotic approximation to the ill-conditioned Riccati equation near the limit provides a robust numerical algorithm by eliminating the large coefficient in the Riccati equation.

The problem is formulated in Section 2 and its solution is derived in Section 3. In Section 4, the limiting properties of the filter are determined. In Section 5, the limiting and asymptotic behaviors of the filter are determined by using the perturbation method. In Section 6, numerical examples are given.

2. Problem formulation

Consider a linear time-varying, uniformly observable system,

$$\dot{x} = Ax + B_u u + B_w w, \tag{1a}$$

$$y = Cx + v, \tag{1b}$$

where u is the control input, y is the measurement, w is the process noise and v is the sensor noise. Following the development in (White & Speyer, 1987; Chung & Speyer, 1998), any plant, actuator and sensor faults can be modeled as additive terms in the state equation (1a). Therefore, a linear system with q faults can be modeled by

$$\dot{x} = Ax + B_u u + B_w w + \sum_{i=1}^{q} \bar{F}_i \bar{\mu}_i, \tag{2a}$$

$$y = Cx + v. \tag{2b}$$

The fault magnitudes $\bar{\mu}_i$ are unknown and arbitrary functions of time that are zero when there is no fault. The fault directions $\bar{F}_i$ are maps that are apriori known. Assume the $\bar{F}_i$'s are monic so that $\bar{\mu}_i \neq 0$ implies $\bar{F}_i \bar{\mu}_i \neq 0$. Since the optimal stochastic fault detection filter is designed to detect only one fault and block other faults, let $\mu_1 = \bar{\mu}_i$ be the target fault and $\mu_2 = [\bar{\mu}_1^{\mathrm{T}} \cdots \bar{\mu}_{i-1}^{\mathrm{T}} \, \bar{\mu}_{i+1}^{\mathrm{T}} \cdots \bar{\mu}_q^{\mathrm{T}}]^{\mathrm{T}}$ be the nuisance fault. Then, (2) can be rewritten as (Massoumnia et al., 1989)

$$\dot{x} = Ax + B_u u + B_w w + F_1 \mu_1 + F_2 \mu_2, \tag{3a}$$

$$y = Cx + v, \tag{3b}$$

where $F_1 = \bar{F}_i$ and $F_2 = [\bar{F}_1 \cdots \bar{F}_{i-1} \, \bar{F}_{i+1} \cdots \bar{F}_q]$.

The objective of the optimal stochastic fault detection filter problem is to find a filter gain L for the linear observer,

$$\dot{\hat{x}} = A\hat{x} + B_u u + L(y - C\hat{x}) \tag{4}$$

and a projector $\hat{H}$ for the residual,

$$r = \hat{H}(y - C\hat{x}) \tag{5}$$

such that the residual is affected primarily by the target fault μ_1 and minimally by the nuisance fault μ_2, process noise w, sensor noise v and initial condition error $x(t_0) - \hat{x}(t_0)$. It is assumed that μ_1, μ_2, w and v are zero mean, white Gaussian noises with power spectral densities Q_1, Q_2, Q_w and V, respectively, and the initial state $x(t_0)$ is a random vector with variance P_0. It is also assumed that μ_1, μ_2, w and v are uncorrelated with each other and with $x(t_0)$.

By using (3) and (4), the dynamic equation of the error, $e = x - \hat{x}$, is

$$\dot{e} = (A - LC)e + F_1 \mu_1 + F_2 \mu_2 + B_w w - Lv. \tag{6}$$

Then, the error can be written as

$$e(t) = \Phi(t, t_0)e(t_0) + \int_{t_0}^{t} \Phi(t, \tau)(F_1 \mu_1 + F_2 \mu_2 + B_w w - Lv)\, \mathrm{d}\tau \tag{7}$$

subject to

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t, t_0) = (A - LC)\Phi(t, t_0), \tag{8}$$

where $\Phi(t_0, t_0) = I$. The residual (5) can be written as $r = \hat{H}(Ce + v)$.

An optimal stochastic fault detection filter problem formulated with a cost criterion based on the residual is unusable from the statistical viewpoint since the variance of the residual generates a δ-function due to the sensor noise. Therefore, the cost criterion will be based on the projected output error $\hat{H}Ce$. In order to determine the cost criterion, define

$$h_1(t) \triangleq \hat{H}C \int_{t_0}^{t} \Phi(t, \tau)F_1 \mu_1 \, \mathrm{d}\tau, \tag{9a}$$

$$h_2(t) \triangleq \hat{H}C \int_{t_0}^{t} \Phi(t, \tau)F_2 \mu_2 \, \mathrm{d}\tau, \tag{9b}$$

$$h_3(t) \triangleq \hat{H}C \left[\Phi(t, t_0)e(t_0) + \int_{t_0}^{t} \Phi(t, \tau)(B_w w - Lv)\, \mathrm{d}\tau \right]. \tag{9c}$$

From (7), $\mathrm{E}[h_1(t)h_1(t)^{\mathrm{T}}]$ represents the transmission from μ_1 to $\hat{H}Ce$, $\mathrm{E}[h_2(t)h_2(t)^{\mathrm{T}}]$ represents the transmission from μ_2 to $\hat{H}Ce$ and $\mathrm{E}[h_3(t)h_3(t)^{\mathrm{T}}]$ represents the transmission from w, v and $e(t_0)$ to $\hat{H}Ce$ where $\mathrm{E}[\bullet]$ is the expectation operator. Note that $e(t_0)$ is a zero mean random vector with variance P_0 if $\hat{x}(t_0) = \mathrm{E}[x(t_0)]$.

The optimal stochastic fault detection filter problem is to find the filter gain L and the projector $\hat{H}$ which minimize the cost criterion,

$$J = \mathrm{tr}\left\{\frac{1}{\gamma}\mathrm{E}[h_2(t)h_2(t)^{\mathrm{T}}] + \mathrm{E}[h_3(t)h_3(t)^{\mathrm{T}}] - \mathrm{E}[h_1(t)h_1(t)^{\mathrm{T}}]\right\}, \tag{10}$$

where t is the current time and γ is a positive scalar. Making γ small places a large weighting on reducing the nuisance fault transmission. The trace operator forms a scalar cost criterion of the matrix output error variance. Note that the power spectral densities Q_1 and Q_2 are considered as design parameters. Since no assumption is made on the fault magnitudes, their white noise representation is a convenience. When Q_1 increases, the transmission from the target fault increases. When Q_2 increases, the transmission from the nuisance fault decreases. However, the power spectral densities Q_w and V, and the variance P_0 can have physical values. When Q_w, V and P_0 increase, the transmission from the process noise, sensor noise and initial condition error decreases, respectively.

Since the effect of the process and sensor noises on the residual is explicitly minimized, the filter is robust with respect to these disturbances. Certain types of model uncertainties can also be modeled as additive noises (Patton & Chen, 1992; Douglas et al., 1997). Therefore, the filter can be made robust to these model uncertainties. In Section 4, it is shown that the filter recovers the geometric structure of the unknown input observer in the limit as $\gamma \to 0$ and the nuisance fault is completely blocked. When it is not at the limit, the filter is an approximate unknown input observer and the nuisance fault is partially blocked. Since the approximate unknown input observer (Chung & Speyer, 1998; Chen & Speyer, 2000) has the additional design freedom to determine how much of the nuisance fault is to be blocked, it is potentially more robust than the classical unknown input observer (Frank, 1990; Massoumnia et al., 1989; Patton & Chen, 1992).

3. Solution

In this section, the minimization problem given by (10) is solved. By using (9), the cost criterion rewritten as

$$J = \mathrm{tr}\left\{\hat{H}C\left[\int_{t_0}^{t}\Phi(t,\tau)\left(LVL^{\mathrm{T}} + \frac{1}{\gamma}F_2Q_2F_2^{\mathrm{T}} - F_1Q_1F_1^{\mathrm{T}} + B_wQ_wB_w^{\mathrm{T}}\right)\Phi(t,\tau)^{\mathrm{T}}\,\mathrm{d}\tau + \Phi(t,t_0)P_0\Phi(t,t_0)^{\mathrm{T}}\right]C^{\mathrm{T}}\hat{H}\right\}$$

is to be minimized with respect to L and $\hat{H}$ subject to (8) and that $\hat{H}$ is a projector. By adding the zero term

$$\mathrm{tr}\left\{\hat{H}C\left[\Phi(t,t)P(t)\Phi(t,t)^{\mathrm{T}} - \Phi(t,t_0)P(t_0)\Phi(t,t_0)^{\mathrm{T}} - \int_{t_0}^{t}\frac{\mathrm{d}}{\mathrm{d}\tau}[\Phi(t,\tau)P(\tau)\Phi(t,\tau)]\,\mathrm{d}\tau\right]C^{\mathrm{T}}\hat{H}\right\}$$

to J and using (8), the minimization problem can be rewritten as

$$\min_{L,\hat{H}}\mathrm{tr}\left[\hat{H}C\int_{t_0}^{t}\Phi(t,\tau)(L - PC^{\mathrm{T}}V^{-1})V(L - PC^{\mathrm{T}}V^{-1})^{\mathrm{T}}\Phi(t,\tau)^{\mathrm{T}}\,\mathrm{d}\tau C^{\mathrm{T}}\hat{H} + \hat{H}CP(t)C^{\mathrm{T}}\hat{H}\right] \tag{11}$$

subject to (8) and that $\hat{H}$ is a projector where

$$\dot{P} = AP + PA^{\mathrm{T}} - PC^{\mathrm{T}}V^{-1}CP + \frac{1}{\gamma}F_2Q_2F_2^{\mathrm{T}} - F_1Q_1F_1^{\mathrm{T}} + B_wQ_wB_w^{\mathrm{T}} \tag{12}$$

and $P(t_0) = P_0$. By inspection, the optimal filter gain is

$$L^* = PC^{\mathrm{T}}V^{-1}. \tag{13}$$

Since $\hat{H}$ is a projector, it can be written as $\hat{H} = \rho\rho^{\mathrm{T}}$ where $\dim\rho = \mathrm{rank}\,\hat{H}$ and $\rho^{\mathrm{T}}\rho = I$. By applying (11) to (13) and substituting $\hat{H} = \rho\rho^{\mathrm{T}}$, the minimization problem reduces to

$$\min_{\rho}\mathrm{tr}[\rho^{\mathrm{T}}CP(t)C^{\mathrm{T}}\rho]$$

subject to $\rho^{\mathrm{T}}\rho = I$. By using a matrix Lagrange multiplier λ to adjoin the constraint to the cost criterion, the first-order necessary condition is obtained as Athans (1968)

$$CP(t)C^{\mathrm{T}}\rho = \rho\lambda.$$

Let $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$ be the eigenvalues of $CP(t)C^{\mathrm{T}}$ and $\rho_1, \rho_2, \ldots, \rho_m$ be the associated eigenvectors. The solution for the optimal ρ depends on the rank of $\hat{H}$. If the rank is chosen as one, the optimal ρ is ρ_m and the optimal projector is

$$\hat{H}^* = \rho_m\rho_m^{\mathrm{T}}. \tag{14}$$

The minimal cost associated with (14) is λ_m. Note that the null space of (14) is $\mathrm{Im}[\rho_1\ \rho_2\cdots\rho_{m-1}]$ because (14) can be written as $\hat{H}^* = I - [\rho_1\ \rho_2\cdots\rho_{m-1}]\,[\rho_1\ \rho_2\cdots\rho_{m-1}]^{\mathrm{T}}$.

In Sections 4 and 5, it is shown that $CP(t)C^{\mathrm{T}}$ has p_2 infinite eigenvalues in the limit as $\gamma \to 0$ and p_2 large eigenvalues near the limit when γ is small where $p_2 = \dim F_2$. Since the remaining $m - p_2$ eigenvalues are very small compared to the p_2 large eigenvalues when γ is small, the rank of $\hat{H}$

can be chosen as $m - p_2$ and the optimal projector is

$$\hat{H}^* = [\rho_m \quad \rho_{m-1} \quad \cdots \quad \rho_{p_2+1}] \quad [\rho_m \quad \rho_{m-1} \quad \cdots \quad \rho_{p_2+1}]^{\mathrm{T}}. \tag{15}$$

The minimal cost associated with (15) is $\sum_{i=p_2+1}^{m} \lambda_i$. The null space of (15) is $\mathrm{Im}[\rho_1\ \rho_2\ \cdots\ \rho_{p_2}]$. Note that both (14) and (15) are optimal projectors depending on the rank chosen. In Sections 4 and 5, it is shown that $\mathrm{Im}[\rho_1\ \rho_2\ \cdots\ \rho_{p_2}]$ contains the nuisance fault completely in the limit and partially near the limit. Thus, the null space of $\hat{H}^*$ only needs to include $\mathrm{Im}[\rho_1\ \rho_2\ \cdots\ \rho_{p_2}]$ in order to block the nuisance fault. Furthermore, (15) allows at least as much of the target fault to pass through as (14) because $\mathrm{Im}[\rho_1\ \rho_2\ \cdots\ \rho_{p_2}] \subseteq \mathrm{Im}[\rho_1\ \rho_2\ \cdots\ \rho_{m-1}]$. Therefore, (15) is a better choice than (14). In Section 4, it is shown that (15) becomes equivalent to the projector used by the unknown input observer in the limit.

Remark 1. To implement the optimal stochastic fault detection filter, the filter gain (13) and the projector (15) are constructed continuously with respect to time because in the cost criterion, t is the current time.

Remark 2. When $Q_1 = 0$, the Riccati matrix P is positive definite. When Q_1 increases, P may become indefinite (Chen, 2000). If Q_1 continues to increase, P may have a finite escape time and goes to $-\infty$. This can be shown by formulating a linear quadratic regulator problem as the dual problem of the optimal stochastic fault detection filter problem and using the result in Speyer (1986). This can be interpreted as an attempt to make the residual sensitive to the target fault. If Q_1 is too large, the target fault may destabilize the filter. Therefore, Q_1 has to be chosen small enough to avoid the finite escape time.

4. Limiting case

In this section, the limiting properties of the optimal stochastic fault detection filter are determined when $\gamma \to 0$. It is shown that the filter places the nuisance fault into an invariant subspace. For time-invariant systems, this invariant subspace is the minimal (C,A)-unobservability subspace of F_2. Therefore, the filter becomes equivalent to the unknown input observer in the limit. For time-varying systems, there exists a similar invariant subspace. Therefore, the filter extends the unknown input observer to the time-varying case. In Section 4.1, the geometric structure of the unknown input observer is given (Massoumnia et al., 1989; Chung & Speyer, 1998). In Section 4.2, the limiting properties of the filter are determined. In Section 4.3, the nuisance fault directions are generalized for time-invariant systems to prevent the invariant zeros of the nuisance fault or their mirror images from becoming part of the eigenvalues of the filter. In Section 4.4, the conditions to ensure that the target fault can be detected are discussed.

4.1. Geometric structure of unknown input observer

The unknown input observer places the nuisance fault into the invariant subspace $\mathscr{T}_2$ which is unobservable to the residual (Massoumnia et al., 1989). $\mathscr{T}_2 = \mathscr{W}_2 \oplus \mathscr{V}_2$ is called the minimal (C,A)-unobservability subspace or the detection space of F_2 (Massoumnia, 1986). $\mathscr{W}_2$ is the minimal (C,A)-invariant subspace of F_2 given by

$$\mathscr{W}_2 = [f_1\ Af_1 \cdots A^{\delta_1} f_1 \quad f_2\ Af_2 \cdots A^{\delta_2} f_2 \cdots f_{p_2}\ Af_{p_2} \cdots A^{\delta_{p_2}} f_{p_2}], \tag{16}$$

where f_i is the ith column of F_2, δ_i is the smallest non-negative integer such that $CA^{\delta_i} f_i \neq 0$ and $p_2 = \dim F_2$. $\mathscr{V}_2$ is the subspace spanned by the invariant zero directions of (C,A,F_2). Note that $\mathscr{T}_2$ is the unobservable subspace of $(\tilde{H}C, A - LC)$ where L is the unknown input observer gain and $\tilde{H}$ is a projector with $\ker \tilde{H} = \mathrm{Im}[CA^{\delta_1} f_1\ CA^{\delta_2} f_2 \cdots CA^{\delta_{p_2}} f_{p_2}]$ (Massoumnia et al., 1989). Therefore, the nuisance fault is unobservable to the residual that uses $\tilde{H}$ as the projector.

For time-varying systems, the minimal (C,A)-invariant subspace of F_2 is (Chung & Speyer, 1998)

$$\mathscr{W}_2 = [b_{1,0}\ b_{1,1} \cdots b_{1,\delta_1} \quad b_{2,0}\ b_{2,1} \cdots b_{2,\delta_2} \cdots b_{p_2,0}\ b_{p_2,1} \cdots b_{p_2,\delta_{p_2}}]. \tag{17}$$

The vectors $b_{i,j}$, $j = 0, 1 \cdots \delta_i$, are obtained from the iteration defined by the Goh transformation, i.e., $b_{i,j} = Ab_{i,j-1} - \dot{b}_{i,j-1}$ with $b_{i,0} = f_i$ where f_i is the ith column of F_2 (Bell & Jacobson, 1975; Moylan & Moore, 1971). δ_i is the smallest non-negative integer such that $Cb_{i,\delta_i} \neq 0$. For time-varying systems, the minimal (C,A)-unobservability subspace cannot be determined because the concept of invariant zero is for time-invariant systems only. The time-varying extension of $\tilde{H}$ is $\ker \tilde{H} = \mathrm{Im}[Cb_{1,\delta_1}\ Cb_{2,\delta_2} \cdots Cb_{p_2,\delta_{p_2}}]$ (Chung & Speyer, 1998).

Remark 3. Eqs. (16) and (17) produce the correct invariant subspaces only when $\mathrm{rank}\, C\mathscr{W}_2 = p_2$. If $\mathrm{rank}\, C\mathscr{W}_2 < p_2$, a new basis for F_2 can be obtained such that $\mathrm{rank}\, C\mathscr{W}_2 = p_2$ (Chen, 2000; Chen & Speyer, 2002).

4.2. Limiting property

In this section, it is assumed that the Riccati matrix P is positive definite. From Remark 2, there always exists

positive definite P for some Q_1. Then, P can be written as

$$P = \sum_{i=1}^{n} \bar{\lambda}_i^{-1} \bar{\rho}_i \bar{\rho}_i^{\mathrm{T}},$$

where $\bar{\lambda}_i^{-1}$ is the ith eigenvalue of P and $\bar{\rho}_i$ is the associated eigenvector. In the limit as $\gamma \to 0$, P goes to infinity because of the term $(1/\gamma)F_2Q_2F_2^{\mathrm{T}}$ in (12) which indicates that some $\bar{\lambda}_i$'s go to zero. Define

$$\Pi \triangleq P^{-1} = \sum_{i=1}^{n} \bar{\lambda}_i \bar{\rho}_i \bar{\rho}_i^{\mathrm{T}}.$$

Then, P goes to infinity in the limit along the null space of Π. By using

$$-\frac{\mathrm{d}}{\mathrm{d}\tau}(P^{-1}) = P^{-1}\left(\frac{\mathrm{d}}{\mathrm{d}\tau}P\right)P^{-1}$$

and (12),

$$-\dot{\Pi} = \Pi A + A^{\mathrm{T}}\Pi + \Pi\left(\frac{1}{\gamma}F_2Q_2F_2^{\mathrm{T}} - F_1Q_1F_1^{\mathrm{T}} + B_wQ_wB_w^{\mathrm{T}}\right)\Pi - C^{\mathrm{T}}V^{-1}C, \tag{18}$$

where $\Pi(t_0) = P_0^{-1}$. Define

$$\bar{\Pi} \triangleq \lim_{\gamma \to 0} \Pi.$$

In the limit, in order for (18) to have a solution,

$$\bar{\Pi}F_2 = 0. \tag{19}$$

This indicates that $\bar{\Pi}$ has a null space which includes F_2. It turns out that $\ker \bar{\Pi}$ is the key to blocking the nuisance fault. Theorem 4 shows that $\ker \bar{\Pi}$ is a (C,A)-invariant subspace. Therefore, the optimal stochastic fault detection filter places the nuisance fault into an invariant subspace in the limit. Theorem 5 shows that $\ker \bar{\Pi}$ also includes the minimal (C,A)-invariant subspace of F_2.

Theorem 4. $\ker \bar{\Pi}$ *is a* (C,A)*-invariant subspace.*

Proof. When only the nuisance fault occurs, the dynamic equation of the error (6) can be written as

$$\Pi\dot{e} = (\Pi A - C^{\mathrm{T}}V^{-1}C)e + \Pi F_2\mu_2.$$

By adding $\dot{\Pi}e$ to both sides and using (18),

$$\frac{\mathrm{d}}{\mathrm{d}\tau}(\Pi e) = -\left[A^{\mathrm{T}} + \Pi\left(\frac{1}{\gamma}F_2Q_2F_2^{\mathrm{T}} - F_1Q_1F_1^{\mathrm{T}} + B_wQ_wB_w^{\mathrm{T}}\right)\right]\Pi e + \Pi F_2\mu_2. \tag{20}$$

In the limit, if the error initially lies in $\ker \bar{\Pi}$, (20) implies that the error will never leave $\ker \bar{\Pi}$ because of (19). Therefore, $\ker \bar{\Pi}$ is a (C,A)-invariant subspace.

Theorem 5. $\ker \bar{\Pi}$ *includes the minimal* (C,A)*-invariant subspace of* F_2.

Proof. Consider the time-varying case first where $\mathscr{W}_2$ is given by (17). From (19), $\bar{\Pi}b_{1,0} = 0$ and $\bar{\Pi}\dot{b}_{1,0} = -\dot{\bar{\Pi}}b_{1,0}$. In the limit, by multiplying (18) by $b_{1,0}^{\mathrm{T}}$ from the left and $b_{1,0}$ from the right, and using $\bar{\Pi}b_{1,0} = 0$,

$$\frac{1}{\gamma}\bar{\Pi}F_2Q_2F_2^{\mathrm{T}}\bar{\Pi}b_{1,0} = 0. \tag{21}$$

By using $\bar{\Pi}\dot{b}_{1,0} = -\dot{\bar{\Pi}}b_{1,0}$, (18) and (21),

$$\bar{\Pi}b_{1,1} = \bar{\Pi}(Ab_{1,0} - \dot{b}_{1,0}) = C^{\mathrm{T}}V^{-1}Cb_{1,0} = 0.$$

From $\bar{\Pi}b_{1,1} = 0$, it can be shown similarly that $\bar{\Pi}b_{1,2} = 0$. By iterating this procedure, $\bar{\Pi}[b_{1,3}\ b_{1,4}\ \cdots\ b_{1,\delta_1}] = 0$. It can be shown similarly that $\bar{\Pi}[b_{i,0}\ b_{i,1}\ \cdots\ b_{i,\delta_i}] = 0$ for $i = 2,3,\ldots,p_2$. Therefore, $\bar{\Pi}\mathscr{W}_2 = 0$. For the time-invariant case, it can be shown similarly.

Whether $\ker \bar{\Pi}$ includes the invariant zero directions of (C,A,F_2) for time-invariant systems is considered now. If $\ker \bar{\Pi}$ does not include the invariant zero directions, the invariant zeros will become part of the filter eigenvalues (i.e., the eigenvalues of $A - LC$) (Massoumnia, 1986). By using the result in Kwakernaak (1976), if there exist left-half-plane invariant zeros, part of the filter eigenvalues will be at the invariant zeros in the limit. If there exist right-half-plane invariant zeros, part of the filter eigenvalues will be at the mirror images of the invariant zeros in the limit. Therefore, $\ker \bar{\Pi}$ includes the invariant zero directions associated with the right-half-plane invariant zeros, but not necessarily the invariant zero directions associated with the left-half-plane invariant zeros. In Section 4.3, the nuisance fault directions are generalized such that $\ker \bar{\Pi}$ includes all the invariant zero directions. This generalization prevents the invariant zeros or their mirror images from becoming part of the filter eigenvalues. This is important because the invariant zeros or their mirror images might be ill-conditioned even though they are in the left-half plane.

For time-invariant systems, $\ker \bar{\Pi} \supseteq \mathscr{W}_2$ from Theorem 5 and $\ker \bar{\Pi} \supseteq \mathscr{V}_2$ from the generalization of the nuisance fault directions. Thus, $\ker \bar{\Pi} \supseteq \mathscr{T}_2$. By using the result in Chung and Speyer (1998) and Chen and Speyer (2000), $\ker \bar{\Pi} \subseteq \mathscr{T}_2$. Therefore, $\ker \bar{\Pi}$ is equivalent to the minimal (C,A)-unobservability subspace of F_2 and the optimal stochastic fault detection filter becomes equivalent to the unknown input observer in the limit. For time-varying systems, $\ker \bar{\Pi} \supseteq \mathscr{W}_2$ from Theorem 5. By using the result in Chen and Speyer (2000), $\ker \bar{\Pi}$ is in the unobservable subspace of $(\tilde{H}C, A - LC)$. Therefore, the optimal stochastic fault detection filter places the nuisance fault into a similar invariant subspace in the limit and extends the unknown input observer to the time-varying case.

Remark 6. By using the optimal filter gain (13) and optimal projector (15), the minimization problem (10) can be written as

$$\frac{\mathrm{tr}\{\mathrm{E}[h_2(t)h_2(t)^{\mathrm{T}}]\} + \gamma\,\mathrm{tr}\{\mathrm{E}[h_3(t)h_3(t)^{\mathrm{T}}]\}}{\mathrm{tr}\{\mathrm{E}[h_1(t)h_1(t)^{\mathrm{T}}]\}} = \gamma\left\{1 + \frac{\sum_{i=p_2+1}^{m}\lambda_i}{\mathrm{tr}\{\mathrm{E}[h_1(t)h_1(t)^{\mathrm{T}}]\}}\right\}.$$

In the limit as $\gamma \to 0$,

$$\frac{\mathrm{tr}\{\mathrm{E}[h_2(t)h_2(t)^{\mathrm{T}}]\}}{\mathrm{tr}\{\mathrm{E}[h_1(t)h_1(t)^{\mathrm{T}}]\}} \to 0$$

This implies that the nuisance fault transmission is zero in the limit.

Remark 7. Since P goes to infinity in the limit along $\ker\bar{\Pi}$, CPC^{T} goes to infinity along $C\ker\bar{\Pi}$. For time-invariant systems, $C\ker\bar{\Pi} = \mathrm{Im}[CA^{\delta_1}f_1\ CA^{\delta_2}f_2\cdots CA^{\delta_{p_2}}f_{p_2}]$. For time-varying systems, $C\ker\bar{\Pi}=\mathrm{Im}[Cb_{1,\delta_1}\ Cb_{2,\delta_2}\cdots Cb_{p_2,\delta_{p_2}}]$. Then, CPC^{T} has p_2 infinite eigenvalues in the limit and their associated eigenvectors span $C\ker\bar{\Pi}$. Therefore, the optimal projector (15) becomes equivalent to $\bar{H}$, which is used by the unknown input observer, in the limit.

4.3. Generalization of nuisance fault direction

The invariant zero of (C,A,F_2) is defined as z at which

$$\begin{bmatrix} zI-A & F_2 \\ C & 0 \end{bmatrix}$$

loses rank. The invariant zero direction v is formed from a partitioning of the null space as

$$\begin{bmatrix} zI-A & F_2 \\ C & 0 \end{bmatrix}\begin{bmatrix} v \\ \bar{v} \end{bmatrix} = 0. \tag{22}$$

From Section 4.2, when f_i, a column vector of F_2, has a left-half-plane invariant zero z_i, $\ker\bar{\Pi}$ includes $\mathrm{Im}[f_i\ Af_i\ \cdots\ A^{\delta_i}f_i]$, but not $\mathrm{Im}\,v_i$ where v_i is the invariant zero direction. Also, z_i becomes one of the filter eigenvalues in the limit. If the nuisance fault direction f_i is replaced by v_i, z_i will not become one of the filter eigenvalues. Furthermore, since $\ker\bar{\Pi}$ includes $\mathrm{Im}[v_i\ Av_i\ \cdots\ A^{\delta_i+1}v_i]$ which is equivalent to $\mathrm{Im}[f_i\ Af_i\ \cdots\ A^{\delta_i}f_i\ v_i]$ by using (22), this generalization will still block the nuisance fault. Note that $\ker\bar{\Pi}$ includes the invariant zero direction now. If the invariant zero is in the right-half plane, this generalization prevents the mirror image of the invariant zero from becoming one of the filter eigenvalues in the limit. If (C,A,v_i) has invariant zeros, the same procedure can be repeated. If the invariant zero is associated with not just one, but several column vectors of F_2, only one of these vectors needs to be replaced by the invariant zero direction.

4.4. Condition on target fault detection

In this section, two conditions to ensure that the target fault can be detected are assumed. First, F_1 and $\ker\bar{\Pi}$ are independent, i.e., $F_1\cap\ker\bar{\Pi}=\emptyset$. Otherwise, the target fault will be difficult or impossible to detect because it will be blocked from the residual along with the nuisance fault even though the filter can still be derived by solving the minimization problem. This condition is similar to but less restrictive than the output separability condition in Massoumnia et al. (1989) and Chung and Speyer (1998), i.e., $C\mathscr{W}_1\cap C\mathscr{W}_2=\emptyset$ where $\mathscr{W}_1$ is the minimal (C,A)-invariant subspace of F_1 which can be obtained similarly by using (16) or (17). The output separability condition is more restrictive because there is an invariant subspace formed for the target fault.

For time-invariant systems, to further ensure a nonzero residual in steady state when the target fault occurs, (C,A,F_1) cannot have invariant zeros at the origin. When only the target fault occurs, the dynamic equation of the error (6) and the residual without the projector can be written as

$$\dot{e} = (A-LC)e + F_1\mu_1,$$

$$r = Ce.$$

For a bias target fault, the residual is zero in steady state if $(C,A-LC,F_1)$ has an invariant zero at the origin (Chen, 1984). Since the filter gain L does not change the invariant zero, $(C,A-LC,F_1)$ has an invariant zero at the origin if and only if (C,A,F_1) has an invariant zero at the origin.

5. Perturbation analysis

In Section 4.2, the limiting properties of the Riccati matrices Π and P were determined. In this section, expressions for Π and P in the limit and near the limit are developed using a perturbation method. The asymptotic expansions of Π and P, explicitly expressed as functions of γ, give an understanding of Π and P when γ is small which is the region of interest for the filter design. In Chen and Speyer (2000) and Chung and Speyer (1998), the Goh transformation in singular optimal control theory (Bell & Jacobson, 1975; Moylan & Moore, 1971) is used to determine Π in the limit. However, the Goh transformation cannot determine Π near the limit. In Section 5.1, Π is expanded around $\gamma=0$. This shows explicitly the characteristics of Π near and in the limit. It is shown that the limiting Π determined from the perturbation method is the same as that determined from the Goh transformation. In Section 5.2, the inverse of Π is derived. This shows explicitly the characteristics of P near and in the limit. The limiting result is consistent with and generalizes the result of Kwakernaak and Sivan (1972).

5.1. Asymptotic expansion

In this section, Π is expanded around $\gamma = 0$ as

$$\Pi = \sum_{i=0}^{\infty} \gamma^{\frac{i}{4}} \Pi_i. \tag{23}$$

By substituting (23) into (18) and collecting terms of common power, the equations used for finding the Π_i's in (23) are obtained in Lemma 8.

Lemma 8.

$$\Pi = [u_1 \quad u_2] \left(\begin{bmatrix} 0 & 0 \\ 0 & \Pi_{022} \end{bmatrix} + \gamma^{1/4} \begin{bmatrix} 0 & 0 \\ 0 & \Pi_{122} \end{bmatrix} + \gamma^{1/2} \begin{bmatrix} \Pi_{211} & \Pi_{212} \\ \Pi_{212}^{\mathrm{T}} & \Pi_{222} \end{bmatrix} + \gamma^{3/4} \begin{bmatrix} \Pi_{311} & \Pi_{312} \\ \Pi_{312}^{\mathrm{T}} & \Pi_{322} \end{bmatrix} + \cdots \right) \times \begin{bmatrix} u_1^{\mathrm{T}} \\ u_2^{\mathrm{T}} \end{bmatrix},$$

where

$$F_2 Q_2 F_2^{\mathrm{T}} = [u_1 \quad u_2] \begin{bmatrix} \sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_1^{\mathrm{T}} \\ u_2^{\mathrm{T}} \end{bmatrix} = u_1 \sigma u_1^{\mathrm{T}}$$

$\sigma > 0$ and $[u_1\ u_2]$ is unitary. Note that $\operatorname{Im} u_1 = \operatorname{Im} F_2$. *$\Pi_{022}$, Π_{211} and Π_{212} must satisfy*:

$$0 = \Pi_{211}\sigma\Pi_{211} - R_{11}, \tag{24a}$$

$$0 = \Pi_{211}\sigma\Pi_{212} + A_{21}^{\mathrm{T}}\Pi_{022} - R_{12}, \tag{24b}$$

$$-\dot{\Pi}_{022} = \Pi_{022}A_{22} + A_{22}^{\mathrm{T}}\Pi_{022} - \Pi_{022}Q_{22}\Pi_{022} - R_{22} + \Pi_{212}^{\mathrm{T}}\sigma\Pi_{212}, \tag{24c}$$

Π_{122}, Π_{311} and Π_{312} must satisfy:

$$0 = \Pi_{311}\sigma\Pi_{211} + \Pi_{211}\sigma\Pi_{311}, \tag{25a}$$

$$0 = \Pi_{311}\sigma\Pi_{212} + \Pi_{211}\sigma\Pi_{312} + A_{21}^{\mathrm{T}}\Pi_{122}, \tag{25b}$$

$$-\dot{\Pi}_{122} = \Pi_{122}(A_{22} - Q_{22}\Pi_{022}) + (A_{22} - Q_{22}\Pi_{022})^{\mathrm{T}}\Pi_{122} + \Pi_{312}^{\mathrm{T}}\sigma\Pi_{212} + \Pi_{212}^{\mathrm{T}}\sigma\Pi_{312}, \tag{25c}$$

where

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \triangleq \begin{bmatrix} u_1^{\mathrm{T}} \\ u_2^{\mathrm{T}} \end{bmatrix} A [u_1 \quad u_2] - \begin{bmatrix} u_1^{\mathrm{T}} \\ u_2^{\mathrm{T}} \end{bmatrix} [\dot{u}_1 \quad \dot{u}_2]$$

$$\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12}^{\mathrm{T}} & Q_{22} \end{bmatrix} \triangleq \begin{bmatrix} u_1^{\mathrm{T}} \\ u_2^{\mathrm{T}} \end{bmatrix} (F_1 Q_1 F_1^{\mathrm{T}} - B_w Q_w B_w^{\mathrm{T}}) \quad [u_1 \quad u_2]$$

$$\begin{bmatrix} R_{11} & R_{12} \\ R_{12}^{\mathrm{T}} & R_{22} \end{bmatrix} \triangleq \begin{bmatrix} u_1^{\mathrm{T}} \\ u_2^{\mathrm{T}} \end{bmatrix} C^{\mathrm{T}} V^{-1} C [u_1 \quad u_2]$$

The equations for the higher-order terms can be found in Chen (2000).

Proof. See Appendix A. □

In Lemma 9, the solution of (24) and (25) is discussed when $CF_2 \neq 0$. In Lemma 10, the solution is discussed when $CF_2 = 0$ and $C(AF_2 - \dot{F}_2) \neq 0$. The higher-order cases, such as $CF_2 = C(AF_2 - \dot{F}_2) = 0$ and $C[A(AF_2 - \dot{F}_2) - \mathrm{d}/\mathrm{d}\tau(AF_2 - \dot{F}_2)] \neq 0$, can be considered similarly.

Lemma 9. *When $CF_2 \neq 0$,*

$$\Pi = [u_1 \quad u_2] \left(\begin{bmatrix} 0 & 0 \\ 0 & \Pi_{022} \end{bmatrix} + \gamma^{1/2} \begin{bmatrix} \Pi_{211} & \Pi_{212} \\ \Pi_{212}^{\mathrm{T}} & \Pi_{212}^{\mathrm{T}}\Pi_{211}^{-1}\Pi_{212} + \bar{\Pi}_{222} \end{bmatrix} + \gamma \cdots \right) \times \begin{bmatrix} u_1^{\mathrm{T}} \\ u_2^{\mathrm{T}} \end{bmatrix}, \tag{26}$$

where

$$\begin{aligned} -\dot{\Pi}_{022} &= \Pi_{022}(A_{22} - A_{21}R_{11}^{-1}R_{12}) + (A_{22} - A_{21}R_{11}^{-1}R_{12})^{\mathrm{T}}\Pi_{022} \\ &\quad + \Pi_{022}(A_{21}R_{11}^{-1}A_{21}^{\mathrm{T}} - Q_{22})\Pi_{022} \\ &\quad - (R_{22} - R_{12}^{\mathrm{T}}R_{11}^{-1}R_{12}), \end{aligned} \tag{27a}$$

$$\Pi_{211} = R_{11}^{1/2}(R_{11}^{1/2}\sigma R_{11}^{1/2})^{-1/2}R_{11}^{1/2}, \tag{27b}$$

$$\Pi_{212} = \sigma^{-1}\Pi_{211}^{-1}(R_{12} - A_{21}^{\mathrm{T}}\Pi_{022}), \tag{27c}$$

$$\dot{\bar{\Pi}}_{222} = \bar{\Pi}_{222}(-A_{22} + Q_{22}\Pi_{022} + A_{21}\Pi_{211}^{-1}\Pi_{212}) + (-A_{22} + Q_{22}\Pi_{022} + A_{21}\Pi_{211}^{-1}\Pi_{212})^{\mathrm{T}}\bar{\Pi}_{222}. \tag{27d}$$

Proof. See Appendix B. □

Lemma 10. *When $CF_2 = 0$ and $C(AF_2 - \dot{F}_2) \neq 0$,*

$$\Pi = [u_1 \quad u_2 v_1 \quad u_2 v_2] \begin{bmatrix} \gamma^{3/4}\Pi_{311} + \gamma \cdots & \gamma^{1/2}\Pi_{2121} + \gamma^{3/4} \cdots & \gamma^{1/2}\Pi_{2122} + \gamma^{3/4} \cdots \\ \gamma^{1/2}\Pi_{2121}^{\mathrm{T}} + \gamma^{3/4} \cdots & \gamma^{1/4}\Pi_{12211} + \gamma^{1/2} \cdots & \gamma^{1/4}\Pi_{12212} + \gamma^{1/2} \cdots \\ \gamma^{1/2}\Pi_{2122}^{\mathrm{T}} + \gamma^{3/4} \cdots & \gamma^{1/4}\Pi_{12212}^{\mathrm{T}} + \gamma^{1/2} \cdots & \Pi_{02222} + \gamma^{1/4} \cdots \end{bmatrix} \begin{bmatrix} u_1^{\mathrm{T}} \\ v_1^{\mathrm{T}}u_2^{\mathrm{T}} \\ v_2^{\mathrm{T}}u_2^{\mathrm{T}} \end{bmatrix}, \tag{28}$$

where $\operatorname{Im} v_1 = \operatorname{Im} A_{21}$ *and* $[v_1\ v_2]$ *is unitary. Only the lowest-order term of each element is kept for simplicity. The equation for each term can be found in Appendix* C *and* Chen (2000).

Proof. See Appendix C. □

When $CF_2 \neq 0$, from Lemma 9,

$$\Pi = [u_1 \quad u_2] \begin{bmatrix} 0 & 0 \\ 0 & \Pi_{022} \end{bmatrix} \begin{bmatrix} u_1^{\mathrm{T}} \\ u_2^{\mathrm{T}} \end{bmatrix} \tag{29}$$

in the limit. Therefore, $\ker \Pi \supseteq \operatorname{Im} u_1 = \operatorname{Im} F_2 = \mathscr{W}_2$ which is consistent with Theorem 5.

Since the Riccati equation (18) can also be generated by solving a differential game similar to the one in Chen and Speyer (2000), the result of (26) gives insight into the singular differential games in Chung and Speyer (1998) and Chen and Speyer (2000). A singular differential game similar to the one in Chen and Speyer (2000) is formulated and solved by using the Goh transformation to derive the limit of (18) when $CF_2 \neq 0$ which is

$$-\dot{S} = S\bar{A} + \bar{A}^{\mathrm{T}}S + S[B_1(F_2^{\mathrm{T}}C^{\mathrm{T}}V^{-1}CF_2)^{-1}B_1^{\mathrm{T}} - F_1Q_1F_1^{\mathrm{T}} + B_wQ_wB_w^{\mathrm{T}}]S - C^{\mathrm{T}}\bar{H}^{\mathrm{T}}V^{-1}\bar{H}C, \tag{30}$$

where $\bar{A} = A - B_1(F_2^{\mathrm{T}}C^{\mathrm{T}}V^{-1}CF_2)^{-1}F_2^{\mathrm{T}}C^{\mathrm{T}}V^{-1}C$, $B_1 = AF_2 - \dot{F}_2$ and $\bar{H} = I - CF_2(F_2^{\mathrm{T}}C^{\mathrm{T}}V^{-1}CF_2)^{-1}F_2^{\mathrm{T}}C^{\mathrm{T}}V^{-1}$. Theorem 11 shows that the limiting Riccati matrix determined from the perturbation method is the same as that determined from the Goh transformation.

Theorem 11.

$$[u_1 \quad u_2] \begin{bmatrix} 0 & 0 \\ 0 & \Pi_{022} \end{bmatrix} \begin{bmatrix} u_1^{\mathrm{T}} \\ u_2^{\mathrm{T}} \end{bmatrix} = S.$$

Proof. See Appendix D. □

When $CF_2 = 0$ and $C(AF_2 - \dot{F}_2) \neq 0$, from Lemma 10,

$$\Pi = [u_1 \quad u_2v_1 \quad u_2v_2] \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Pi_{02222} \end{bmatrix} \begin{bmatrix} u_1^{\mathrm{T}} \\ v_1^{\mathrm{T}}u_2^{\mathrm{T}} \\ v_2^{\mathrm{T}}u_2^{\mathrm{T}} \end{bmatrix} \tag{31}$$

in the limit. By using $\operatorname{Im} v_1 = \operatorname{Im} A_{21}$ and $\operatorname{Im} u_1 = \operatorname{Im} F_2$,

$$\begin{aligned} \operatorname{Im}[u_1 \quad u_2v_1] &= \operatorname{Im}[u_1 \quad u_2u_2^{\mathrm{T}}(Au_1 - \dot{u}_1)] \\ &= \operatorname{Im}[u_1 \quad (I - u_1u_1^{\mathrm{T}})(Au_1 - \dot{u}_1)] \\ &= \operatorname{Im}[u_1 \quad Au_1 - \dot{u}_1] = \operatorname{Im}[F_2 \quad AF_2 - \dot{F}_2] \end{aligned}$$

Therefore, $\ker \Pi \supseteq \operatorname{Im}[F_2\ AF_2 - \dot{F}_2] = \mathscr{W}_2$ which is consistent with Theorem 5.

5.2. Analysis

In this section, an expression for the inverse of Π is derived. This shows explicitly the characteristics of P near and in the limit. Only time-invariant systems are considered because Π_{022} in (29) and Π_{02222} in (31) may not be invertible for time-varying systems. In Lemma 12, P is determined when $CF_2 \neq 0$. In Lemma 13, P is determined when $CF_2 = 0$ and $CAF_2 \neq 0$. The higher-order cases, such as $CF_2 = CAF_2 = 0$ and $CA^2F_2 \neq 0$, can be considered similarly.

Lemma 12. *When* $CF_2 \neq 0$,

$$P = [u_1 \quad u_2] \left(\gamma^{-1/2} \begin{bmatrix} \Pi_{211}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \Pi_{211}^{-1}(\Pi_{212}\Pi_{022}^{-1}\Pi_{212}^{\mathrm{T}} - \Pi_{411})\Pi_{211}^{-1} & -\Pi_{211}^{-1}\Pi_{212}\Pi_{022}^{-1} \\ -\Pi_{022}^{-1}\Pi_{212}^{\mathrm{T}}\Pi_{211}^{-1} & \Pi_{022}^{-1} \end{bmatrix} + \gamma^{1/2}\cdots \right) \begin{bmatrix} u_1^{\mathrm{T}} \\ u_2^{\mathrm{T}} \end{bmatrix}. \tag{32}$$

Proof. By using Lemma 9 and matrix inversion lemma, (32) is obtained.

Lemma 13. *When* $CF_2 = 0$ *and* $CAF_2 \neq 0$,

$$P = [u_1 \quad u_2v_1 \quad u_2v_2] \begin{bmatrix} \gamma^{-3/4}P_{11} + \gamma^{-1/2}\cdots & \gamma^{-1/2}P_{12} + \gamma^{-1/4}\cdots & P_{13} + \gamma^{1/4}\cdots \\ \gamma^{-1/2}P_{12}^{\mathrm{T}} + \gamma^{-1/4}\cdots & \gamma^{-1/4}P_{22} + \cdots & P_{23} + \gamma^{1/4}\cdots \\ P_{13}^{\mathrm{T}} + \gamma^{1/4}\cdots & P_{23}^{\mathrm{T}} + \gamma^{1/4}\cdots & P_{33} + \gamma^{1/4}\cdots \end{bmatrix} \begin{bmatrix} u_1^{\mathrm{T}} \\ v_1^{\mathrm{T}}u_2^{\mathrm{T}} \\ v_2^{\mathrm{T}}u_2^{\mathrm{T}} \end{bmatrix}, \tag{33}$$

where P_{ij}, $i,j = 1,\ldots,3$, *can be found in* Chen (2000). *Only the lowest-order term of each element is kept for simplicity.*

Proof. By using Lemma 10 and matrix inversion lemma, (33) is obtained (Chen, 2000).

In the limit, when $CF_2 \neq 0$, Lemma 12 shows that P goes to infinity along the direction of $\operatorname{Im} F_2$. In the limit, when $CF_2 = 0$ and $CAF_2 \neq 0$, Lemma 13 shows that P goes to infinity along the direction of $\operatorname{Im}[F_2\ AF_2]$.

Remark 14. By using the result in Kwakernaak and Sivan (1972), for the time-invariant and infinite-time case, under the assumption that (C, A, F_2) does not have right-half-plane invariant zeros,

$$\gamma P \to 0 \tag{34a}$$

$$L \to \gamma^{-1/2}F_2Q_2^{1/2}U^{\mathrm{T}}V^{-1/2} \tag{34b}$$

as $\gamma \to 0$ where U is an arbitrary matrix such that $U^{\mathrm{T}}U = I$.

To compare this result with Lemmas 12 and 13, (34a) is satisfied by multiplying (32) and (33) by γ. By substituting (32) into (13),

$$L \to \gamma^{-1/2} u_1 \Pi_{211}^{-1} u_1^{\mathrm{T}} C^{\mathrm{T}} V^{-1}$$

as $\gamma \to 0$. Therefore, L goes to infinity along the direction of $\gamma^{-1/2}\mathrm{Im}\, F_2$ which is consistent with (34b). By substituting (33) into (13),

$$L \to \gamma^{-1/2} u_1 P_{12} v_1^{\mathrm{T}} u_2^{\mathrm{T}} C^{\mathrm{T}} V^{-1} + \gamma^{-1/4} u_2 v_1 P_{22} v_1^{\mathrm{T}} u_2^{\mathrm{T}} C^{\mathrm{T}} V^{-1}$$

as $\gamma \to 0$. Therefore, L goes to infinity essentially along the direction of $\gamma^{-1/2}\mathrm{Im}\, F_2$ which is consistent with (34b). However, L also goes to infinity along the direction of $\gamma^{-1/4}\mathrm{Im}\, u_2 v_1$ where $\mathrm{Im}\,[F_2\ u_2 v_1] = \mathrm{Im}\,[F_2\ AF_2]$. Therefore, the perturbation method is consistent with and generalizes the result of Kwakernaak and Sivan (1972).

6. Example

In this section, two numerical examples are used to demonstrate the performance of the optimal stochastic fault detection filter. In Section 6.1, the filter is applied to a time-invariant system. In Section 6.2, the filter is applied to a time-varying system.

6.1. Example 1

Consider the time-invariant system from White and Speyer (1987),

$$A = \begin{bmatrix} 0 & 3 & 4 \\ 1 & 2 & 3 \\ 0 & 2 & 5 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$F_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad F_2 = \begin{bmatrix} 5 \\ 1 \\ 1 \end{bmatrix},$$

where F_1 is the target fault direction and F_2 is the nuisance fault direction. There is no process noise. To determine the optimal stochastic fault detection filter, the power spectral densities are chosen as $Q_1 = 1$, $Q_2 = 1$ and $V = I$. The steady-state solutions to the Riccati equation (12) when $\gamma = 10^{-4}$ and 10^{-6} are obtained, respectively. Fig. 1 shows the frequency response from both faults to residual (5). The left one is obtained with $\gamma = 10^{-4}$ and the right one is obtained with $\gamma = 10^{-6}$. In each figure, there are two solid lines representing the frequency response from the target fault to the residuals using projectors (15) and $\tilde{H}$, respectively. Note that these two solid lines overlap. The dashdot line and dashed line represent the frequency response from the nuisance fault to the residuals using projectors (15)

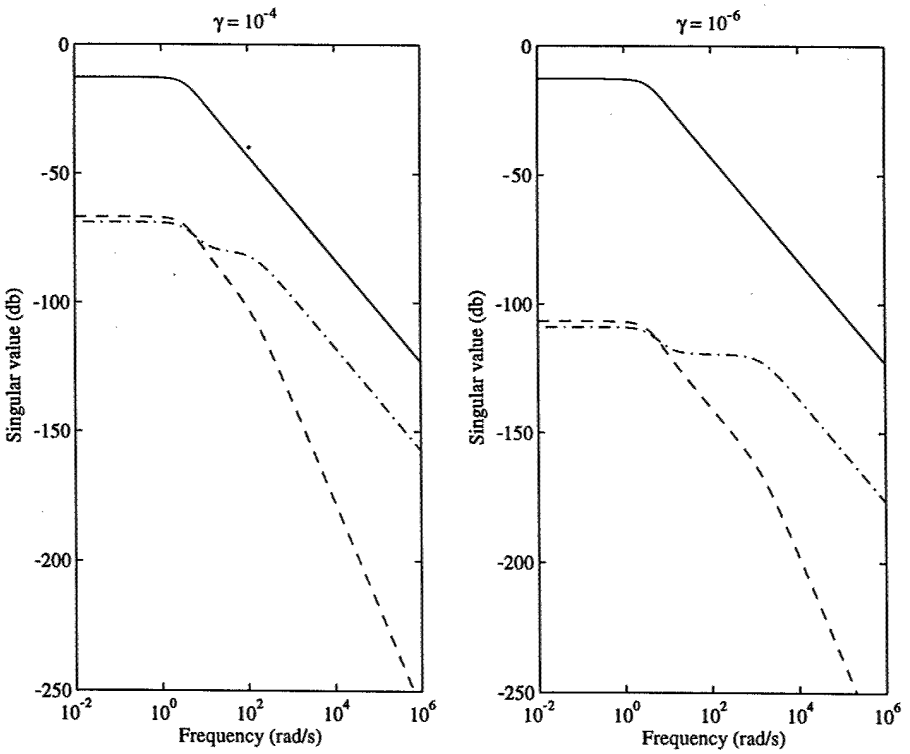


Fig. 1. Frequency response from both faults to the residual.

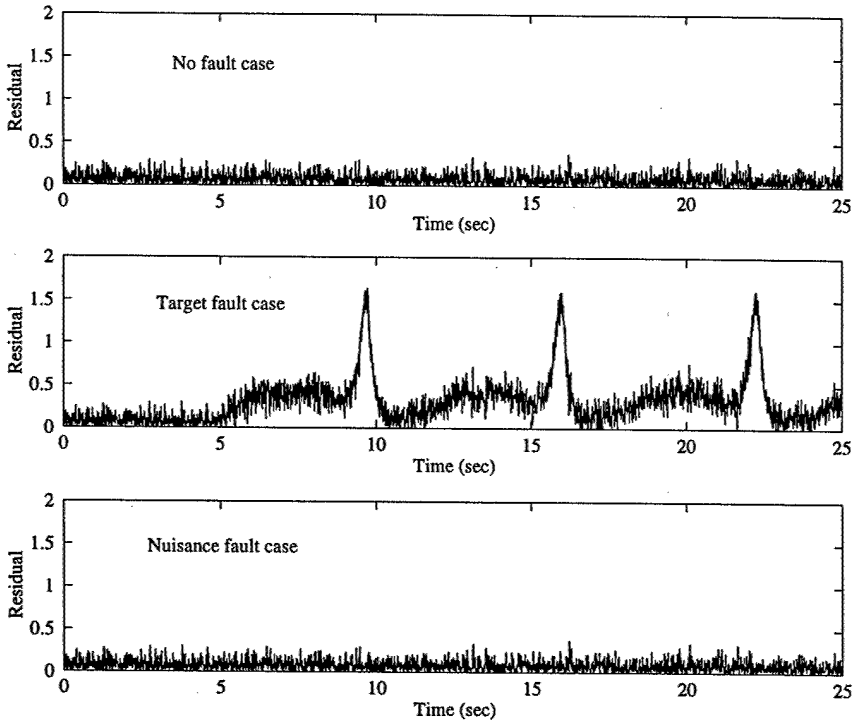


Fig. 2. Time response of the residual.

and $\tilde{H}$, respectively. This example shows that the nuisance fault transmission can be reduced by using a smaller γ while the target fault transmission remains large. Furthermore, the projector (15), derived from solving the minimization problem, is slightly better than $\tilde{H}$, the projector used by other approximate unknown input observers (Chung & Speyer, 1998; Chen & Speyer, 2000), at low frequency. This suggests that $\tilde{H}$ might not be the best choice for the approximate unknown input observer.

6.2. Example 2

Consider a time-varying system obtained by adding some time-varying elements to the time-invariant system in previous section,

$$A = \begin{bmatrix} -\cos(t) & 3+2\sin(t) & 4 \\ 1 & 2 & 3-2\cos(t) \\ 5\sin(t) & 2 & 5+3\cos(t) \end{bmatrix},$$

$$F_2 = \begin{bmatrix} 5-2\cos(t) \\ 1 \\ 1+\sin(t) \end{bmatrix}$$

while C and F_1 remain the same. The Riccati equation (12) is solved with $Q_1 = 1$, $Q_2 = 1$, $V = I$, $P_0 = I$ and $\gamma = 10^{-4}$ for $t \in [0,25]$. Fig. 2 shows the time response of the norm of residual (5) using projector (15) when there is no fault, a target fault and a nuisance fault, respectively. The faults are steps of magnitude 3 that occur at the fifth second. The sensor noise is a zero mean, white Gaussian noise with power spectral density of $10^{-4}I$. This example shows that the residual is very sensitive to the target fault and much less sensitive to the nuisance fault. Therefore, the filter performs well for time-varying systems.

7. Conclusion

The optimal stochastic fault detection filter is derived from solving a stochastic minimization problem. In the limit, the filter recovers the geometric structure of the unknown input observer and the nuisance fault is completely blocked. When it is not at the limit, the filter is an approximate unknown input observer and the nuisance fault is partially blocked. The perturbation method used to obtain the limiting and asymptotic behaviors of the filter can be applied to other approximate unknown input observers (Chung & Speyer, 1998; Chen & Speyer, 2000) derived by solving differential games which consider the worst-case scenarios. For time-invariant systems, the filter performance can be enhanced by replacing the nuisance fault directions with the invariant zero directions. This notion can also be applied to other approximate unknown input observers. Finally, filter designs can be obtained for both time-invariant and time-varying systems.

Acknowledgements

This work was sponsored by Air Force Office of Scientific Research F49620-00-1-0154, NASA Goddard Space Flight Center NAG5-11384 and California Department of Transportation TO 4209.

Appendix A. Proof of Lemma 8

By substituting (23) into (18) and collecting terms of common power,

$$\gamma^{-1} : 0 = \Pi_0\bar{Q}_2\Pi_0, \tag{A.1a}$$

$$\gamma^{-3/4} : 0 = \Pi_1\bar{Q}_2\Pi_0 + \Pi_0\bar{Q}_2\Pi_1, \tag{A.1b}$$

$$\gamma^{-1/2} : 0 = \Pi_2\bar{Q}_2\Pi_0 + \Pi_1\bar{Q}_2\Pi_1 + \Pi_0\bar{Q}_2\Pi_2, \tag{A.1c}$$

$$\gamma^{-1/4} : 0 = \Pi_3\bar{Q}_2\Pi_0 + \Pi_2\bar{Q}_2\Pi_1 + \Pi_1\bar{Q}_2\Pi_2 + \Pi_0\bar{Q}_2\Pi_3, \tag{A.1d}$$

$$\gamma^{0} : -\dot{\Pi}_0 = \Pi_0A + A^{\mathrm{T}}\Pi_0 - C^{\mathrm{T}}V^{-1}C + \Pi_4\bar{Q}_2\Pi_0 + \Pi_3\bar{Q}_2\Pi_1 + \Pi_2\bar{Q}_2\Pi_2 + \Pi_1\bar{Q}_2\Pi_3 + \Pi_0\bar{Q}_2\Pi_4 - \Pi_0\bar{Q}_1\Pi_0, \tag{A.1e}$$

$$\gamma^{1/4} : -\dot{\Pi}_1 = \Pi_1A + A^{\mathrm{T}}\Pi_1 + \Pi_5\bar{Q}_2\Pi_0 + \Pi_4\bar{Q}_2\Pi_1 + \Pi_3\bar{Q}_2\Pi_2 + \Pi_2\bar{Q}_2\Pi_3 + \Pi_1\bar{Q}_2\Pi_4 + \Pi_0\bar{Q}_2\Pi_5 - \Pi_1\bar{Q}_1\Pi_0 - \Pi_0\bar{Q}_1\Pi_1, \tag{A.1f}$$

where $\bar{Q}_2 = F_2Q_2F_2^{\mathrm{T}}$ and $\bar{Q}_1 = F_1Q_1F_1^{\mathrm{T}} - B_wQ_wB_w^{\mathrm{T}}$.

From (A.1a), Π_0 can be written as

$$\Pi_0 = [u_1 \quad u_2]\begin{bmatrix}0 & 0\\ 0 & \Pi_{022}\end{bmatrix}\begin{bmatrix}u_1^{\mathrm{T}}\\ u_2^{\mathrm{T}}\end{bmatrix} = u_2\Pi_{022}u_2^{\mathrm{T}}, \tag{A.2}$$

where Π_{022} is to be determined. (A.1b) is trivially satisfied because of (A.2). By substituting (A.2) into (A.1c), Π_1 can be written as

$$\Pi_1 = [u_1 \quad u_2]\begin{bmatrix}0 & 0\\ 0 & \Pi_{122}\end{bmatrix}\begin{bmatrix}u_1^{\mathrm{T}}\\ u_2^{\mathrm{T}}\end{bmatrix} = u_2\Pi_{122}u_2^{\mathrm{T}}, \tag{A.3}$$

where Π_{122} is to be determined. (A.1d) is trivially satisfied because of (A.2) and (A.3).

Let

$$\Pi_2 = [u_1 \quad u_2]\begin{bmatrix}\Pi_{211} & \Pi_{212}\\ \Pi_{212}^{\mathrm{T}} & \Pi_{222}\end{bmatrix}\begin{bmatrix}u_1^{\mathrm{T}}\\ u_2^{\mathrm{T}}\end{bmatrix}. \tag{A.4}$$

By multiplying (A.1e) by $[u_1\ u_2]^{\mathrm{T}}$ from the left and $[u_1\ u_2]$ from the right, and substituting (A.2), (A.3) and (A.4), (24) is obtained. Let

$$\Pi_3 = [u_1 \quad u_2]\begin{bmatrix}\Pi_{311} & \Pi_{312}\\ \Pi_{312}^{\mathrm{T}} & \Pi_{322}\end{bmatrix}\begin{bmatrix}u_1^{\mathrm{T}}\\ u_2^{\mathrm{T}}\end{bmatrix}. \tag{A.5}$$

By multiplying (A.1f) by $[u_1\ u_2]^{\mathrm{T}}$ from the left and $[u_1\ u_2]$ from the right, and substituting (A.2), (A.3), (A.4) and (A.5), (25) is obtained. The same procedure can be used to obtain the equations for the higher-order terms if needed (Chen, 2000; Chen, Mingori, & Speyer, 2001).

Remark 15. Since $\mathrm{Im}\,u_1 = \mathrm{Im}\,F_2$, u_1 can be chosen as $F_2(F_2^{\mathrm{T}}F_2)^{-1/2}$. Since $[u_1\ u_2]$ is unitary, u_2 has to satisfy $u_1^{\mathrm{T}}u_2 = 0$ and $u_2^{\mathrm{T}}u_2 = I$. Define $U_1 = I - u_1u_1^{\mathrm{T}}$. Since $u_1^{\mathrm{T}}U_1 = 0$, the first column of u_2, called u_{21}, can be chosen as $u_{21} = U_{1i}(U_{1i}^{\mathrm{T}}U_{1i})^{-1/2}$ where U_{1i} is any nonzero column of U_1. Next, define $U_2 = I - [u_1\ u_{21}][u_1\ u_{21}]^{\mathrm{T}}$. Then, the second column of u_2, called u_{22}, can be chosen as $u_{22} = U_{2i}(U_{2i}^{\mathrm{T}}U_{2i})^{-1/2}$ where U_{2i} is any nonzero column of U_2. Other directions of u_2 can be obtained similarly. $\dot{u}_1$ and $\dot{u}_2$ can also be obtained since u_1 and u_2 are explicitly written as functions of time. For time-invariant systems, $[u_1\ u_2]$ can also be obtained from the singular value decomposition of $F_2Q_2F_2^{\mathrm{T}}$ and $[\dot{u}_1\ \dot{u}_2] = 0$. Note that $[u_1\ u_2]$ is generally not unique. However, the theorem and all lemmas in Section 5 are true for any $[u_1\ u_2]$ satisfying $\mathrm{Im}\,u_1 = \mathrm{Im}\,F_2$ and $[u_1\ u_2]$ is unitary.

Appendix B. Proof of Lemma 9

When $CF_2 \neq 0$, R_{11} is positive definite because $\mathrm{Im}\,u_1 = \mathrm{Im}\,F_2$. Then, from (24a), (27b) is obtained. Note that Π_{211} is positive definite. From (24b), (27c) is obtained. By substituting (27c) into (24c) and using (24a), (27a) is obtained. Therefore, the zeroth-order term Π_0 (A.2) can be obtained from (27a). Part of the second-order term Π_2 (A.4) can be obtained from (27b) and (27c).

From (25a), $\Pi_{311} = 0$ because σ and Π_{211} are positive definite. By substituting $\Pi_{311} = 0$ into (25b),

$$\Pi_{312} = -\sigma^{-1}\Pi_{211}^{-1}A_{21}^{\mathrm{T}}\Pi_{122}. \tag{B.1}$$

By substituting (B.1) into (25c),

$$\dot{\Pi}_{122} = \Pi_{122}(-A_{22} + Q_{22}\Pi_{022} + A_{21}\Pi_{211}^{-1}\Pi_{212}) + (-A_{22} + Q_{22}\Pi_{022} + A_{21}\Pi_{211}^{-1}\Pi_{212})^{\mathrm{T}}\Pi_{122}. \tag{B.2}$$

Since (B.2) is a homogeneous equation and the initial condition is zero, $\Pi_{122} = 0$. By substituting $\Pi_{122} = 0$ into (B.1), $\Pi_{312} = 0$. Therefore, the first-order term Π_1 (A.3) and part of the third-order term Π_3 (A.5) are zero. Similar procedure can be used to obtain (27d) (Chen, 2000; Chen et al., 2001). Therefore, the second-order term Π_2 (A.4) can be obtained from (27b), (27c) and (27d). Similar procedure can be used to obtain the equations for the higher-order terms if needed (Chen, 2000; Chen et al., 2001). It can be shown that the rest of the odd terms (i.e., $\Pi_3, \Pi_5, \ldots$) are zero. Therefore, when $CF_2 \neq 0$, the expansion of Π (23) only needs to be in the order of $\gamma^{1/2}$.

Remark 16. Since Π_{211} and Π_{212} are obtained from algebraic equations (27b) and (27c), the initial condition $\Pi(t_0)=P_0^{-1}$ cannot be satisfied in general. This is because the dimension of the Riccati equation (18) is reduced in the limit as $\gamma \to 0$ which leads to the occurrence of a boundary layer (Nayfeh, 1973). The expansion of Π (26) is called the outer expansion and valid everywhere except near $\tau=0$. The inner expansion, which is valid only near $\tau=0$, can be obtained by using different fast time scales (Nayfeh, 1973). Since the inner expansion is only valid for a very short period of time, only the boundary layer is obtained and used as the initial condition of the outer expansion (Chen et al., 2001). Note that in the limit, the fast time scale goes to infinity and there is an instant jump at the initial time which is consistent with the Goh transformation (Chen & Speyer, 2000; Chung & Speyer, 1998).

Appendix C. Proof of Lemma 10

When $CF_2=0$, $R_{11}=0$ and $R_{12}=0$ because $\operatorname{Im} u_1=\operatorname{Im} F_2$. From (24a), $\Pi_{211}=0$ because σ is positive definite. By substituting $\Pi_{211}=0$ into (24b), $\Pi_{022}A_{21}=0$. Then, Π_{022} can be written as

$$\Pi_{022}=[v_1 \quad v_2]\begin{bmatrix}0 & 0\\ 0 & \Pi_{02222}\end{bmatrix}\begin{bmatrix}v_1^{\mathrm{T}}\\ v_2^{\mathrm{T}}\end{bmatrix}. \tag{C.1}$$

Let

$$\Pi_{212}=[\Pi_{2121} \quad \Pi_{2122}]\begin{bmatrix}v_1^{\mathrm{T}}\\ v_2^{\mathrm{T}}\end{bmatrix}. \tag{C.2}$$

$$\Pi=[u_1 \quad u_2]\begin{bmatrix}\gamma^{3/4}\Pi_{311}+\gamma\cdots & \gamma^{1/2}\Pi_{212}+\gamma^{3/4}\cdots\\ \gamma^{1/2}\Pi_{212}^{\mathrm{T}}+\gamma^{3/4}\cdots & [v_1 \quad v_2]\begin{bmatrix}\gamma^{1/4}\Pi_{12211}+\gamma^{1/2}\cdots & \gamma^{1/4}\Pi_{12212}+\gamma^{1/2}\cdots\\ \gamma^{1/4}\Pi_{12212}^{\mathrm{T}}+\gamma^{1/2}\cdots & \Pi_{02222}+\gamma^{1/4}\cdots\end{bmatrix}\begin{bmatrix}v_1^{\mathrm{T}}\\ v_2^{\mathrm{T}}\end{bmatrix}\end{bmatrix}\begin{bmatrix}u_1^{\mathrm{T}}\\ u_2^{\mathrm{T}}\end{bmatrix}$$

By multiplying (24c) by $[v_1\ v_2]^{\mathrm{T}}$ from the left and $[v_1\ v_2]$ from the right, and substituting (C.1) and (C.2),

$$0=\Pi_{2121}^{\mathrm{T}}\sigma\Pi_{2121}-R_{2211}, \tag{C.3a}$$

$$0=\Pi_{2121}^{\mathrm{T}}\sigma\Pi_{2122}+A_{2221}^{\mathrm{T}}\Pi_{02222}-R_{2212}, \tag{C.3b}$$

$$-\dot{\Pi}_{02222}=\Pi_{02222}A_{2222}+A_{2222}^{\mathrm{T}}\Pi_{02222}-\Pi_{02222}Q_{2222}\Pi_{02222}-R_{2222}+\Pi_{2122}^{\mathrm{T}}\sigma\Pi_{2122}, \tag{C.3c}$$

where

$$\begin{bmatrix}A_{2211} & A_{2212}\\ A_{2221} & A_{2222}\end{bmatrix}\triangleq\begin{bmatrix}v_1^{\mathrm{T}}\\ v_2^{\mathrm{T}}\end{bmatrix}A_{22}[v_1 \quad v_2]-\begin{bmatrix}v_1^{\mathrm{T}}\\ v_2^{\mathrm{T}}\end{bmatrix}[\dot{v}_1 \quad \dot{v}_2],$$

$$\begin{bmatrix}Q_{2211} & Q_{2212}\\ Q_{2212}^{\mathrm{T}} & Q_{2222}\end{bmatrix}\triangleq\begin{bmatrix}v_1^{\mathrm{T}}\\ v_2^{\mathrm{T}}\end{bmatrix}Q_{22}[v_1 \quad v_2],$$

$$\begin{bmatrix}R_{2211} & R_{2212}\\ R_{2212}^{\mathrm{T}} & R_{2222}\end{bmatrix}\triangleq\begin{bmatrix}v_1^{\mathrm{T}}\\ v_2^{\mathrm{T}}\end{bmatrix}R_{22}[v_1 \quad v_2].$$

Since $\operatorname{Im} u_1=\operatorname{Im} F_2$, $Cu_1=0$ and $C(Au_1-\dot{u}_1)\neq 0$. Since $R_{2211}=v_1^{\mathrm{T}}u_2^{\mathrm{T}}C^{\mathrm{T}}V^{-1}Cu_2v_1$ and $\operatorname{Im} v_1=\operatorname{Im} A_{21}$, R_{2211} is positive definite because

$$\begin{aligned}A_{21}^{\mathrm{T}}u_2^{\mathrm{T}}C^{\mathrm{T}}V^{-1}Cu_2A_{21}&=(u_1^{\mathrm{T}}A^{\mathrm{T}}u_2-\dot{u}_1^{\mathrm{T}}u_2)u_2^{\mathrm{T}}C^{\mathrm{T}}V^{-1}Cu_2(u_2^{\mathrm{T}}Au_1-u_2^{\mathrm{T}}\dot{u}_1)\\&=(u_1^{\mathrm{T}}A^{\mathrm{T}}-\dot{u}_1^{\mathrm{T}})(I-u_1u_1^{\mathrm{T}})C^{\mathrm{T}}V^{-1}C(I-u_1u_1^{\mathrm{T}})(Au_1-\dot{u}_1)\\&=(Au_1-\dot{u}_1)^{\mathrm{T}}C^{\mathrm{T}}V^{-1}C(Au_1-\dot{u}_1)>0\end{aligned}$$

Then, Π_{2121} is invertible. From (C.3b),

$$\Pi_{2122}=\sigma^{-1}\Pi_{2121}^{-T}(R_{2212}-A_{2221}^{\mathrm{T}}\Pi_{02222}). \tag{C.4}$$

By substituting (C.4) into (C.3c) and using (C.3a),

$$\begin{aligned}-\dot{\Pi}_{02222}&=\Pi_{02222}(A_{2222}-A_{2221}R_{2211}^{-1}R_{2212})\\&\quad+(A_{2222}-A_{2221}R_{2211}^{-1}R_{2212})^{\mathrm{T}}\Pi_{02222}\\&\quad+\Pi_{02222}(A_{2221}R_{2211}^{-1}A_{2221}^{\mathrm{T}}-Q_{2222})\Pi_{02222}\\&\quad-(R_{2222}-R_{2212}^{\mathrm{T}}R_{2211}^{-1}R_{2212}).\end{aligned} \tag{C.5}$$

Therefore, the zeroth-order term Π_0 (A.2) can be obtained from (C.1) and (C.5). Similar procedure can be used to obtain the equations for other terms (Chen, 2000; Chen et al., 2001). Therefore, Π can be expressed as

which can be written as (28).

Appendix D. Proof of Theorem 11

Since $SF_2=0$ (Chen & Speyer, 2000), S can be written as

$$S=[u_1 \quad u_2]\begin{bmatrix}0 & 0\\ 0 & \bar{S}\end{bmatrix}\begin{bmatrix}u_1^{\mathrm{T}}\\ u_2^{\mathrm{T}}\end{bmatrix}.$$

By multiplying (30) by $[u_1\ u_2]^{\mathrm{T}}$ from the left and $[u_1\ u_2]$ from the right, subtracting $[u_1\ u_2]^{\mathrm{T}}S[\dot{u}_1\ \dot{u}_2]$ and $[\dot{u}_1\ \dot{u}_2]^{\mathrm{T}}S[u_1\ u_2]$ from both sides, and using $[u_1\ u_2]\,[u_1\ u_2]^{\mathrm{T}}=I$,

$$\begin{bmatrix}0 & 0\\ 0 & -\dot{\bar{S}}\end{bmatrix}=S_1+S_1^{\mathrm{T}}+S_2-S_0, \tag{D.1}$$

where

$$S_1 = \begin{bmatrix} 0 & 0 \\ 0 & \bar{S} \end{bmatrix} \left(\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} - \begin{bmatrix} u_1^{\mathrm{T}} \\ u_2^{\mathrm{T}} \end{bmatrix} B_1 (F_2^{\mathrm{T}} C^{\mathrm{T}} V^{-1} C F_2)^{-1} F_2^{\mathrm{T}} C^{\mathrm{T}} V^{-1} C [u_1 \quad u_2] \right), \quad \text{(D.2a)}$$

$$S_2 = \begin{bmatrix} 0 & 0 \\ 0 & \bar{S} \end{bmatrix} \left(\begin{bmatrix} u_1^{\mathrm{T}} \\ u_2^{\mathrm{T}} \end{bmatrix} B_1 (F_2^{\mathrm{T}} C^{\mathrm{T}} V^{-1} C F_2)^{-1} B_1^{\mathrm{T}} [u_1 \quad u_2] - \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12}^{\mathrm{T}} & Q_{22} \end{bmatrix} \right) \begin{bmatrix} 0 & 0 \\ 0 & \bar{S} \end{bmatrix}, \quad \text{(D.2b)}$$

$$S_0 = \begin{bmatrix} R_{11} & R_{12} \\ R_{12}^{\mathrm{T}} & R_{22} \end{bmatrix} - \begin{bmatrix} R_{11} & R_{12} \\ R_{12}^{\mathrm{T}} & R_{22} \end{bmatrix} \begin{bmatrix} u_1^{\mathrm{T}} \\ u_2^{\mathrm{T}} \end{bmatrix} F_2 (F_2^{\mathrm{T}} C^{\mathrm{T}} V^{-1} C F_2)^{-1} F_2^{\mathrm{T}} [u_1 \quad u_2] \begin{bmatrix} R_{11} & R_{12} \\ R_{12}^{\mathrm{T}} & R_{22} \end{bmatrix}. \quad \text{(D.2c)}$$

Since $\operatorname{Im} u_1 = \operatorname{Im} F_2$, let $u_1 = F_2 \Sigma$ where Σ satisfies $\Sigma^{\mathrm{T}} F_2^{\mathrm{T}} F_2 \Sigma = I$ because $u_1^{\mathrm{T}} u_1 = I$. By using $u_1 = F_2 \Sigma$ and $\Sigma^{-1} = \Sigma^{\mathrm{T}} F_2^{\mathrm{T}} F_2$,

$$u_1^{\mathrm{T}} F_2 (F_2^{\mathrm{T}} C^{\mathrm{T}} V^{-1} C F_2)^{-1} F_2^{\mathrm{T}} u_1 = [\Sigma^{\mathrm{T}} (F_2^{\mathrm{T}} C^{\mathrm{T}} V^{-1} C F_2) \Sigma]^{-1} = R_{11}^{-1}. \quad \text{(D.3)}$$

By using $u_2^{\mathrm{T}} F_2 = 0$ and (D.3),

$$S_0 = \begin{bmatrix} 0 & 0 \\ 0 & R_{22} - R_{12}^{\mathrm{T}} R_{11}^{-1} R_{12} \end{bmatrix}. \quad \text{(D.4)}$$

Since $\dot{u}_1 = \dot{F}_2 \Sigma + F_2 \dot{\Sigma}$, $u_2^{\mathrm{T}} \dot{u}_1 = u_2^{\mathrm{T}} \dot{F}_2 \Sigma$ because $u_2^{\mathrm{T}} F_2 = 0$. By using $B_1 = A F_2 - \dot{F}_2$, $u_1 = F_2 \Sigma$ and $u_2^{\mathrm{T}} \dot{F}_2 \Sigma = u_2^{\mathrm{T}} \dot{u}_1$,

$$u_2^{\mathrm{T}} B_1 \Sigma = A_{21}. \quad \text{(D.5)}$$

By using (D.3), $u_1 = F_2 \Sigma$, $\Sigma^{-1} = \Sigma^{\mathrm{T}} F_2^{\mathrm{T}} F_2$ and (D.5),

$$u_2^{\mathrm{T}} B_1 (F_2^{\mathrm{T}} C^{\mathrm{T}} V^{-1} C F_2)^{-1} B_1^{\mathrm{T}} u_2 = A_{21} R_{11}^{-1} A_{21}^{\mathrm{T}}. \quad \text{(D.6)}$$

By using (D.6),

$$S_2 = \begin{bmatrix} 0 & 0 \\ 0 & \bar{S}(A_{21} R_{11}^{-1} A_{21}^{\mathrm{T}} - Q_{22}) \bar{S} \end{bmatrix}. \quad \text{(D.7)}$$

By using $u_1 = F_2 \Sigma$ and (D.5),

$$u_2^{\mathrm{T}} B_1 (F_2^{\mathrm{T}} C^{\mathrm{T}} V^{-1} C F_2)^{-1} F_2^{\mathrm{T}} C^{\mathrm{T}} V^{-1} C u_1 = A_{21}. \quad \text{(D.8)}$$

By using (D.3), $u_1 = F_2 \Sigma$, $\Sigma^{-1} = \Sigma^{\mathrm{T}} F_2^{\mathrm{T}} F_2$ and (D.5),

$$u_2^{\mathrm{T}} B_1 (F_2^{\mathrm{T}} C^{\mathrm{T}} V^{-1} C F_2)^{-1} F_2^{\mathrm{T}} C^{\mathrm{T}} V^{-1} C u_2 = A_{21} R_{11}^{-1} R_{12}. \quad \text{(D.9)}$$

By using (D.8) and (D.9),

$$S_1 = \begin{bmatrix} 0 & 0 \\ 0 & \bar{S}(A_{22} - A_{21} R_{11}^{-1} R_{12}) \end{bmatrix}. \quad \text{(D.10)}$$

By substituting (D.10), (D.7) and (D.4) into (D.1),

$$-\dot{\bar{S}} = \bar{S}(A_{22} - A_{21} R_{11}^{-1} R_{12}) + (A_{22} - A_{21} R_{11}^{-1} R_{12})^{\mathrm{T}} \bar{S} + \bar{S}(A_{21} R_{11}^{-1} A_{21}^{\mathrm{T}} - Q_{22}) \bar{S} - (R_{22} - R_{12}^{\mathrm{T}} R_{11}^{-1} R_{12}). \quad \text{(D.11)}$$

By comparing (D.11) with (27a), $\bar{S} = \Pi_{022}$.

References

Athans, M. (1968). The matrix minimum principle. *Information and Control, 11*(5/6), 592–606.

Bell, D. J., & Jacobson, D. H. (1975). *Singular optimal control problems*. New York: Academic Press.

Chen, C. -T. (1984). *Linear system theory and design*. New York: Holt, Rinehart, and Winston.

Chen, R. H., 2000. *Fault detection filters for robust analytical redundancy*. Ph.D. thesis, University of California at Los Angeles.

Chen, R. H., Mingori, D. L., & Speyer, J. L. (2001). Perturbation analysis for Ricatti equation. In *Proceedings of the American control conference* (pp. 3463–3468).

Chen, R. H., & Speyer, J. L. (2000). A generalized least-squares fault detection filter. *International Journal of Adaptive Control and Signal Processing—Special Issue: Fault Detection and Isolation, 14*(7), 747–757.

Chen, R. H., & Speyer, J. L. (2002). Robust multiple-fault detection filter. *International Journal of Robust and Nonlinear Control—Special Issue: Fault Detection and Isolation, 12*(8), 675–696.

Chung, W. H., & Speyer, J. L. (1998). A game theoretic fault detection filter. *IEEE Transactions on Automatic Control, AC-43*(2), 143–161.

Douglas, R. K., Chen, R. H., & Speyer, J. L. (1997). Model input reduction. In *Proceedings of the American Control Conference* (pp. 3882–3886).

Frank, P. M. (1990). Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy—a survey and some new results. *Automatica, 26*(3), 459–474.

Kwakernaak, H. (1976). Asymptotic root loci of multivariable linear optimal regulators. *IEEE Transactions on Automatic Control, AC-21*(3), 378–382.

Kwakernaak, H., & Sivan, R. (1972). The maximally achievable accuracy of linear optimal regulators and linear optimal filters. *IEEE Transactions on Automatic Control, AC-17*(1), 79–86.

Massoumnia, M. A. (1986). A geometric approach to the synthesis of failure detection filters. *IEEE Transactions on Automatic Control, AC-31*(9), 839–846.

Massoumnia, M. A., Verghese, G. C., & Willsky, A. S. (1989). Failure detection and identification. *IEEE Transactions on Automatic Control, AC-34*(3), 316–321.

Moylan, P. J., & Moore, J. B. (1971). Generalizations of singular optimal control theory. *Automatica, 7*(5), 591–598.

Nayfeh, A. H. (1973). *Perturbation methods*. New York: Wiley-Interscience.

Patton, R. J., & Chen, J. (1992). Robust fault detection of jet engine sensor systems using eigenstructure assignment. *AIAA Journal of Guidance, Control, and Dynamics, 15*(6), 1491–1497.

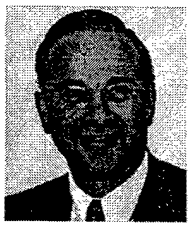
Speyer, J. L. (1986). Linear-quadratic control problem. In C. T. Leondes (Ed.), *Control and dynamic systems*, Vol. 13 (pp. 241–293). New York: Academic Press.

White, J. E., & Speyer, J. L. (1987). Detection filter design: Spectral theory and algorithms. *IEEE Transactions on Automatic Control, AC-32*(7), 593–603.

Robert H. Chen received the B.S. degree in power mechanical engineering from National Tsing Hua University, Taiwan, in 1992 and the Ph.D. degree in mechanical engineering from the University of California, Los Angeles, in 2000.

He is currently an assistant research engineer in the Mechanical and Aerospace Engineering Department at the University of California, Los Angeles. His research interests include fault detection, identification, and reconstruction with application to aircraft, satellites, and ground vehicles.

D. Lewis Mingori received the B.S. degree from the University of California, Berkeley in 1960, the MS degree from UCLA in 1962, and the Ph.D. degree from Stanford University in 1966 (Aeronautical and Astronautical Sciences). From 1966–1968 he was a Member of the Technical Staff of The Aerospace Corporation, and since 1968 he has been a faculty member in the Mechanical and Aerospace Engineering Department at UCLA. His teaching is in the areas of Dynamics and Control and his research interests include Spacecraft Attitude Control, Spinning and Dual Spin Spacecraft Dynamics, Spinning Rocket Dynamics, Automated Road Vehicles, Fault Detection and Identification, and Modeling, Identification and Vibration Control of Flexible Structures. Dr. Mingori is a Fellow of the American Institute of Aeronautics and Astronautics.



Jason L. Speyer received the S.B. degree in aeronautics and astronautics from the Massachusetts Institute of Technology, Cambridge, in 1960 and the Ph.D. degree in applied mathematics from Harvard University, Cambridge, MA, in 1968.

His industrial experience includes research at Boeing, Raytheon, Analytical Mechanics Associated, and the Charles Stark Draper Laboratory. He was the Harry H. Power Professor in Aerospace Engineering at the University of Texas, Austin, and is currently a Professor in the Mechanical and Aerospace Engineering Department at the University of California, Los Angeles. He spent a research leave as a Lady Davis Visiting Professor at the Technion—Israel Institute of Technology, Haifa, Israel, in 1983 and was the 1990 Jerome C. Hunsaker Visiting Professor of Aeronautics and Astronautics at the Massachusetts Institute of Technology.

Dr. Speyer has twice been an elected member of the Board of Governors of the IEEE Control Systems Society. He has served as an Associate Editor of the IEEE Transactions on Automatic Control and as Chairman of the Technical Committee on Aerospace Control. He is a Fellow of the American Institute of Aeronautics and Astronautics and the Institute of Electrical and Electronic Engineers. From October 1987 to October 1991 and from October 1997 to October 2001, he has served as a member of the USAF Scientific Advisory Board. He was awarded Mechanics and Control of Flight Award and Dryden Lectureship in Research from the American Institute of Aeronautics and Astronautics in 1985 and 1995, respectively. He was awarded Air Force Exceptional Civilian Decoration in 1991 and 2001 and the IEEE Third Millennium Medal in 2000.

Appendix H

"Fault Reconstruction from Sensor and Actuator Failures,"

Robert H. Chen and Jason L. Speyer,

Proceedings of the IEEE Conference on Decision and Control, December, 2001

Sensor and Actuator Fault Reconstruction

Robert H. Chen and Jason L. Speyer
Mechanical and Aerospace Engineering Department
University of California, Los Angeles, California 90095-1597
chrobert@talus.seas.ucla.edu, speyer@seas.ucla.edu

Abstract

Many fault detection filters have been developed to detect and identify sensor and actuator faults. An approach to further reconstruct sensor and actuator faults from the residual generated by the fault detection filter is proposed. The transfer matrix from the faults to the residual is derived in terms of the eigenvalues of the fault detection filter associated with the invariant subspaces of the faults and the invariant zeros of the faults. For each fault, all possible fault reconstruction processes are derived and parameterized by applying a projector to the transfer matrix and taking inverse. Then, the optimal fault reconstruction process is determined by minimizing the ratio of the $\mathcal{H}_2$ norm of the projected transfer matrix from the disturbance to the $\mathcal{H}_2$ norm of the projected transfer matrix from the fault. For the existence of the fault reconstruction process, the invariant zeros of the fault have to be in the left-half plane. Furthermore, for reconstructing a sensor fault, the system has to be detectable with respect to the other sensors.

1 Introduction

Any system under automatic control demands a high degree of reliability in order to operate properly. If a sensor fails, the controller's command will be generated using incorrect measurement. If an actuator fails, the controller's command will not be applied properly to the system. Therefore, one needs a health monitoring system capable of detecting a fault as it occurs and identifying the faulty component. This process is called fault detection and identification. One approach to fault detection and identification is the fault detection filter which was first introduced by [1] and refined by [2]. It is also known as Beard-Jones detection filter. A geometric interpretation and a spectral analysis of the fault detection filter are given in [3] and [4], respectively. The idea of the fault detection filter is to place the reachable subspace of each fault into invariant subspaces which do not overlap each other. Then, when a nonzero residual is detected, a fault can be announced and identified by projecting the residual onto each of the invariant subspaces. Design algorithms have been developed to improve the robustness of the fault detection filter [5, 6].

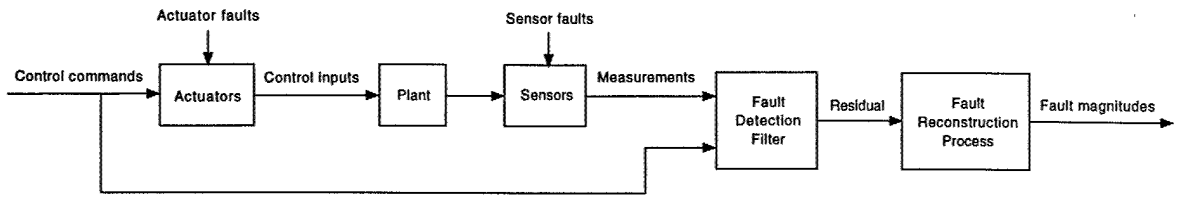


Figure 1: Fault detection, identification and reconstruction

When these faults occur, the residual only has a transient response and becomes zero after a while even though the faults still exist. A bias in a single position sensor is one possible example. However, for some of these faults, the fault reconstruction process can still generate the magnitudes of the faults even after the residual becomes zero.

In Section 2, the background of the fault detection filter is given. In Section 3, the transfer matrix from the fault to the residual is derived. In Section 4, the reconstruction of sensor and actuator faults is discussed. In Section 5, a numerical example is given.

2 Fault Detection Filter Background

In this section, the background of the fault detection filter is given [1, 3, 4, 9]. This is important because the fault reconstruction process uses the residual generated by the fault detection filter to generate the magnitudes of actuator and sensor faults.

Consider a linear time-invariant system,

$$\dot{x} = Ax + Bu \tag{1a}$$

$$y = Cx \tag{1b}$$

where u is the control input and y is the measurement. The ith actuator fault can be modeled as an additive term in the state equation (1a) [1, 4].

$$\dot{x} = Ax + Bu + F_a\mu_a$$

where F_a is the ith column of B and μ_a is an unknown and arbitrary scalar function of time that is zero when there is no fault. The failure mode μ_a models the time-varying amplitude of the actuator fault while the failure signature F_a models the directional characteristics of the actuator fault. The ith sensor fault can be modeled as an additive term in the measurement equation (1b) [1, 4].

$$y = Cx + E_s\mu_s \tag{2}$$

The fault detection filter gain L is chosen such that $A - LC$ is stable and there exists an invariant subspace $\mathcal{T}_i$ for each fault F_i. $\mathcal{T}_i$ is called the minimal (C, A)-unobservability subspace or the detection space of F_i. Assume that the invariant zeros of (C, A, F_i) have the same geometric and algebraic multiplicities. Then, $\mathcal{T}_i$ can be obtained by

$$\mathcal{T}_i = \mathcal{W}_i \oplus \mathcal{V}_i \tag{7}$$

where $\mathcal{W}_i$ is the minimal (C, A)-invariant subspace of F_i given by the recursive algorithm

$$\mathcal{W}_i^{k+1} = \mathrm{Im} F_i \oplus A(\mathcal{W}_i^k \cap \mathrm{Ker}\, C) \quad \text{where} \quad \mathcal{W}_i^0 = \emptyset \tag{8}$$

and $\mathcal{V}_i$ is spanned by the invariant zero directions of (C, A, F_i). The invariant zero of (C, A, F_i) is defined as z at which $\begin{bmatrix} A - zI & F_i \\ C & 0 \end{bmatrix}$ loses rank. The invariant zero direction ν is formed from a partitioning of the null space as

$$\begin{bmatrix} A - zI & F_i \\ C & 0 \end{bmatrix} \begin{bmatrix} \nu \\ \bar{\nu} \end{bmatrix} = 0 \tag{9}$$

When $\dim F_i = 1$, the recursive algorithm (8) implies

$$\mathcal{W}_i = \mathrm{Im} \begin{bmatrix} F_i & AF_i & \cdots & A^{k_i} F_i \end{bmatrix}$$

where k_i is the smallest non-negative integer such that $CA^{k_i} F_i \neq 0$. It is assumed that $\mathcal{T}_1 \cdots \mathcal{T}_q$ are independent. If they are not independent, the faults can only be detected, but not identified. This condition is called output separability. It is also assumed that $(C, A, [F_1 \cdots F_q])$ does not have more invariant zeros than $(C, A, F_1) \cdots (C, A, F_q)$. If it does, the extra invariant zeros will become part of the eigenvalues of $A - LC$. This condition is called mutual detectability. For more details, please refer to [3]. For the algorithms to form the fault detection filter gain L, please refer to [4, 5, 6].

When there is no fault, the residual is zero after the transient response due to the initial condition error because $A - LC$ is stable. When the fault μ_i occurs, the residual becomes nonzero, but only in the direction of $C\mathcal{T}_i$ because $\mathrm{Im}\, F_i \subseteq \mathcal{T}_i$ and $(A - LC)\mathcal{T}_i \subseteq \mathcal{T}_i$. By using a projector

$$\hat{H}_i = I - \mathrm{Ker}\, \hat{H}_i \left[(\mathrm{Ker}\, \hat{H}_i)^T \mathrm{Ker}\, \hat{H}_i \right]^{-1} (\mathrm{Ker}\, \hat{H}_i)^T \tag{10}$$

where $\mathrm{Ker}\, \hat{H}_i = [C\mathcal{T}_1 \cdots C\mathcal{T}_{i-1}\ C\mathcal{T}_{i+1} \cdots C\mathcal{T}_q]$, the projected residual $\hat{H}_i r$ is only sensitive to the fault μ_i, but not to the other faults $\mu_{j \neq i}$. Therefore, the fault detection filter can detect and identify actuator and sensor faults.

After the faulty sensor or actuator has been detected and identified, the system may switch to an identical redundant sensor or actuator. If such sensor or actuator is not available, the controller has to be reconfigured based on the remaining non-faulty sensors and actuators. However, if the magnitude of the sensor fault can be obtained, the correct measurement can be obtained by subtracting the fault from the faulty measurement. Then, the controller may continue to function normally without the need for reconfiguration. This is particularly useful when an intermittent sensor fault occurs. If the magnitude of the actuator fault can be obtained, the control input applied to the system can be obtained by adding the fault to the control command. Then, the condition of the actuator can be diagnosed and the controller can be reconfigured such that the faulty actuator may be compensated. For example, if a bias is developed in the actuator, it may be compensated by reconfiguring the controller given the size of the bias. If the actuator is stuck in certain position, it can be diagnosed because the control input would be a constant regardless of the control command. Then, the controller may be reconfigured by using the remaining non-faulty actuators to compensate the faulty actuator allowing continued operation of the system. Therefore, fault reconstruction increases the flexibility of the system's reaction to sensor and actuator faults.

In this paper, an approach for reconstructing sensor and actuator faults is presented. The fault reconstruction process generates the magnitudes of sensor and actuator faults using the residual generated by the fault detection filter. The block diagram is shown in Figure 1. The transfer matrix from the faults to the residual is derived in terms of the eigenvalues of the fault detection filter associated with the invariant subspaces of the faults and the invariant zeros of the faults. By applying a projector to the transfer matrix, a projected residual that is only sensitive to one fault, but not to the other faults, is obtained. By taking the inverse of the projected transfer matrix, all possible fault reconstruction processes are derived and parameterized. Then, the optimal fault reconstruction process is determined by minimizing the ratio of the $\mathcal{H}_2$ norm of the transfer matrix from the disturbance to the projected residual over the $\mathcal{H}_2$ norm of the transfer matrix from the fault to the projected residual. For the existence of the fault reconstruction process, the invariant zeros of the fault have to be in the left-half plane. Furthermore, for reconstructing a sensor fault, the system has to be detectable with respect to the other sensors. Note that the fault reconstruction process can also be derived numerically from the state-space models of the plant and fault detection filter by using the Silverman's algorithm [7, 8]. However, the Silverman's algorithm can produce only one particular fault reconstruction process which is not optimal in general. Furthermore, the existence conditions and the analytical structure of the fault reconstruction process cannot be obtained.

In addition to be used for fault reconstruction, the transfer matrix from the faults to the residual provides a frequency domain interpretation of the fault detection filter which complements the geometric interpretation by [3]. The transfer matrix provides information about the transient and steady-state responses of the residual to the faults. It is shown explicitly the types of faults that the fault detection filter cannot detect.

where E_s is a column of zeros except a one in the ith position and μ_s is an unknown and arbitrary scalar function of time that is zero when there is no fault. The failure mode μ_s models the time-varying amplitude of the sensor fault while the failure signature E_s models the directional characteristics of the sensor fault. For the purpose of fault detection filter design, an input to the state equation (1a) which drives the measurement in the same way that μ_s does in (2) is obtained as in [9]. Define a new state $\bar{x} = x + f_s\mu_s$ where $E_s = Cf_s$. Then, the dynamic equation of $\bar{x}$ and (2) can be written as

$$\dot{\bar{x}} = A\bar{x} + Bu + \begin{bmatrix} f_s & \bar{f}_s \end{bmatrix} \begin{bmatrix} \dot{\mu}_s \\ -\mu_s \end{bmatrix} \tag{3a}$$

$$y = C\bar{x} \tag{3b}$$

where $\bar{f}_s = Af_s$. Hence, for fault detection filter design, the sensor fault can be modeled as a two-dimensional additive term in the state equation as in (3).

Therefore, a linear time-invariant system with q_a actuator faults and q_s sensor faults can be modeled as

$$\dot{x} = Ax + Bu + \sum_{i=1}^{q_a} F_{ai}\mu_{ai} \tag{4a}$$

$$y = Cx + \sum_{i=1}^{q_s} E_{si}\mu_{si} \tag{4b}$$

However, for fault detection filter design, the following system is used.

$$\dot{x} = Ax + Bu + \sum_{i=1}^{q_a} F_{ai}\mu_{ai} + \sum_{i=1}^{q_s} \begin{bmatrix} f_{si} & \bar{f}_{si} \end{bmatrix} \begin{bmatrix} \dot{\mu}_{si} \\ -\mu_{si} \end{bmatrix} = Ax + Bu + \sum_{i=1}^{q} F_i\mu_i \tag{5a}$$

$$y = Cx \tag{5b}$$

where $q = q_a + q_s$. For $i = 1 \cdots q_a$, $F_i = F_{ai}$ and $\mu_i = \mu_{ai}$. For $i = 1 \cdots q_s$, $F_{i+q_a} = [\, f_{si} \;\; \bar{f}_{si} \,]$ and $\mu_{i+q_a} = [\, \dot{\mu}_{si} \;\; -\mu_{si} \,]^T$.

Now the fault detection filter will be introduced from the geometric point of view [3]. The design algorithms [4, 5, 6] are omitted because only the geometric properties of the fault detection filter are involved with the derivation of the fault reconstruction process. Assume (C, A) is observable. Fault detection filter is a linear observer in the form of

$$\dot{\hat{x}} = A\hat{x} + Bu + L(y - C\hat{x}) \tag{6a}$$

$$r = y - C\hat{x} \tag{6b}$$

where r is called the residual. By using (5) and (6), the dynamic equation of the error $e = x - \hat{x}$ and the residual can be written as

$$\dot{e} = (A - LC)e + \sum_{i=1}^{q} F_i\mu_i$$

$$r = Ce$$

3 Transfer Matrix from the Fault to the Residual

In this section, the transfer matrix from the fault to the residual is derived in terms of the eigenvalues of the fault detection filter associated with the detection space of the fault and the invariant zeros of the fault. This gives a frequency domain interpretation of the fault detection filter which complements the geometric interpretation by [3]. The transfer matrix provides information about the transient and steady-state responses of the residual to the fault. It is shown explicitly the types of faults that the fault detection filter cannot detect. In Section 3.1, the actuator fault is considered. In Section 3.2, the sensor fault is considered.

3.1 Actuator Fault

From (4) and (6), the transfer matrix from the actuator fault μ_{ai} to the residual r is

$$\frac{r(s)}{\mu_{ai}(s)} = C(sI - A + LC)^{-1}F_{ai}$$

When (C, A, F_{ai}) has p_{ai} invariant zeros at $z_{ai,1} \cdots z_{ai,p_{ai}}$, from (7), the detection space of F_{ai} is

$$\mathcal{T}_{ai} = \mathrm{Im}\left[\begin{array}{cccccccc} F_{ai} & AF_{ai} & \cdots & A^{k_{ai}}F_{ai} & \nu_{ai,1} & \nu_{ai,2} & \cdots & \nu_{ai,p_{ai}} \end{array}\right]$$

where k_{ai} is the smallest non-negative integer such that $CA^{k_{ai}}F_{ai} \neq 0$ and $\nu_{ai,1} \cdots \nu_{ai,p_{ai}}$ are the invariant zero directions. Let $\delta_{ai} \triangleq \dim \mathcal{T}_{ai} = k_{ai} + p_{ai} + 1$. Assume that $\lambda_{ai,1} \cdots \lambda_{ai,\delta_{ai}}$, the eigenvalues of $A - LC$ associated with $\mathcal{T}_{ai}$, are distinct. Since $\mathcal{T}_{ai}$ spans δ_{ai} eigenvectors of $A - LC$,

$$(A - LC)x_j = \lambda_{ai,j}x_j \tag{11}$$

where $j = 1 \cdots \delta_{ai}$ and

$$\left[\begin{array}{cccccccc} F_{ai} & AF_{ai} & \cdots & A^{k_{ai}}F_{ai} & \nu_{ai,1} & \cdots & \nu_{ai,p_{ai}} \end{array}\right] = \left[\begin{array}{cccc} x_1 & x_2 & \cdots & x_{\delta_{ai}} \end{array}\right] \left[\begin{array}{cccc} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,\delta_{ai}} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,\delta_{ai}} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{\delta_{ai},1} & \alpha_{\delta_{ai},2} & \cdots & \alpha_{\delta_{ai},\delta_{ai}} \end{array}\right] \tag{12}$$

If $\lambda_{ai,1} \cdots \lambda_{ai,\delta_{ai}}$ are not distinct, (11) may be modified with the generalized eigenvectors. For $1 \leq k \leq k_{ai}$,

$$\begin{aligned}(A - LC)A^k F_{ai} &= (A - LC)A^k F_{ai} - (A - LC)LCA^{k-1}F_{ai} = (A - LC)^2 A^{k-1}F_{ai} = \cdots \\ &= (A - LC)^{k+1}F_{ai} = (A - LC)^{k+1}\sum_{j=1}^{\delta_{ai}} \alpha_{j,1}x_j = \sum_{j=1}^{\delta_{ai}} \lambda_{ai,j}^{k+1}\alpha_{j,1}x_j\end{aligned}$$

and

$$(A - LC)A^k F_{ai} = (A - LC)\sum_{j=1}^{\delta_{ai}} \alpha_{j,k+1}x_j = \sum_{j=1}^{\delta_{ai}} \lambda_{ai,j}\alpha_{j,k+1}x_j$$

The resulting relationship is

$$\alpha_{j,k+1} = \lambda_{ai,j}^{k} \alpha_{j,1} \tag{13}$$

for $j = 1 \cdots \delta_{ai}$ and $k = 1 \cdots k_{ai}$. For $1 \leq k \leq p_{ai}$, by using (9),

$$(A - LC)\nu_{ai,k} = A\nu_{ai,k} = z_{ai,k}\nu_{ai,k} - F_{ai}\bar{\nu}_{ai,k} = \sum_{j=1}^{\delta_{ai}} (z_{ai,k}\alpha_{j,k+k_{ai}+1} - \bar{\nu}_{ai,k}\alpha_{j,1})x_j$$

and

$$(A - LC)\nu_{ai,k} = (A - LC)\sum_{j=1}^{\delta_{ai}} \alpha_{j,k+k_{ai}+1}x_j = \sum_{j=1}^{\delta_{ai}} \lambda_{ai,j}\alpha_{j,k+k_{ai}+1}x_j$$

Assume that $\lambda_{ai,j} \neq z_{ai,k}$, the resulting relationship is

$$\alpha_{j,k+k_{ai}+1} = \frac{\bar{\nu}_{ai,k}\alpha_{j,1}}{z_{ai,k} - \lambda_{ai,j}} \tag{14}$$

for $j = 1 \cdots \delta_{ai}$ and $k = 1 \cdots p_{ai}$. For the case where $\lambda_{ai,j} = z_{ai,k}$, please see Appendix. By substituting (13) and (14) into (12),

$$\begin{bmatrix} F_{ai} & AF_{ai} & \cdots & A^{k_{ai}}F_{ai} & \nu_{ai,1} & \cdots & \nu_{ai,p_{ai}} \end{bmatrix}$$

$$= \begin{bmatrix} \alpha_{1,1}x_1 & \alpha_{2,1}x_2 & \cdots & \alpha_{\delta_{ai},1}x_{\delta_{ai}} \end{bmatrix} \begin{bmatrix} 1 & \lambda_{ai,1} & \cdots & \lambda_{ai,1}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,1}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,1}} \\ 1 & \lambda_{ai,2} & \cdots & \lambda_{ai,2}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,2}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,2}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{ai,\delta_{ai}} & \cdots & \lambda_{ai,\delta_{ai}}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,\delta_{ai}}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,\delta_{ai}}} \end{bmatrix} \tag{15}$$

From (11),

$$(sI - A + LC)x_j = (s - \lambda_{ai,j})x_j \Rightarrow (sI - A + LC)^{-1}x_j = \frac{1}{s - \lambda_{ai,j}}x_j$$

Then,

$$C(sI - A + LC)^{-1}F_{ai} = \sum_{j=1}^{\delta_{ai}} \frac{\alpha_{j,1}}{s - \lambda_{ai,j}} Cx_j = C \begin{bmatrix} \alpha_{1,1}x_1 & \alpha_{2,1}x_2 & \cdots & \alpha_{\delta_{ai},1}x_{\delta_{ai}} \end{bmatrix} \begin{bmatrix} \frac{1}{s-\lambda_{ai,1}} \\ \frac{1}{s-\lambda_{ai,2}} \\ \vdots \\ \frac{1}{s-\lambda_{ai,\delta_{ai}}} \end{bmatrix} \tag{16}$$

By using (15),

$$C(sI - A + LC)^{-1}F_{ai} = CA^{k_{ai}}F_{ai} \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & \lambda_{ai,1} & \cdots & \lambda_{ai,1}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,1}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,1}} \\ 1 & \lambda_{ai,2} & \cdots & \lambda_{ai,2}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,2}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,2}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{ai,\delta_{ai}} & \cdots & \lambda_{ai,\delta_{ai}}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,\delta_{ai}}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,\delta_{ai}}} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{s-\lambda_{ai,1}} \\ \frac{1}{s-\lambda_{ai,2}} \\ \vdots \\ \frac{1}{s-\lambda_{ai,\delta_{ai}}} \end{bmatrix}$$

By using Cramer's rule in matrix theory [10],

$$C(sI-A+LC)^{-1}F_{ai}$$

$$= \frac{\det\begin{bmatrix} 1 & \lambda_{ai,1} & \cdots & \lambda_{ai,1}^{k_{ai}-1} & \frac{1}{s-\lambda_{ai,1}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,1}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,1}} \\ 1 & \lambda_{ai,2} & \cdots & \lambda_{ai,2}^{k_{ai}-1} & \frac{1}{s-\lambda_{ai,2}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,2}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,2}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{ai,\delta_{ai}} & \cdots & \lambda_{ai,\delta_{ai}}^{k_{ai}-1} & \frac{1}{s-\lambda_{ai,\delta_{ai}}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,\delta_{ai}}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,\delta_{ai}}} \end{bmatrix}}{\det\begin{bmatrix} 1 & \lambda_{ai,1} & \cdots & \lambda_{ai,1}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,1}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,1}} \\ 1 & \lambda_{ai,2} & \cdots & \lambda_{ai,2}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,2}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,2}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{ai,\delta_{ai}} & \cdots & \lambda_{ai,\delta_{ai}}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,\delta_{ai}}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,\delta_{ai}}} \end{bmatrix}} CA^{k_{ai}}F_{ai} \qquad (17)$$

By using the determinant operations in matrix theory [10], the numerator of (17) becomes

$$\frac{1}{\prod_{j=1}^{\delta_{ai}}(s-\lambda_{ai,j})} \det\begin{bmatrix} s-\lambda_{ai,1} & \cdots & (s-\lambda_{ai,1})\lambda_{ai,1}^{k_{ai}-1} & 1 & \frac{\bar{\nu}_{ai,1}(s-\lambda_{ai,1})}{z_{ai,1}-\lambda_{ai,1}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}(s-\lambda_{ai,1})}{z_{ai,p_{ai}}-\lambda_{ai,1}} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ s-\lambda_{ai,\delta_{ai}} & \cdots & (s-\lambda_{ai,\delta_{ai}})\lambda_{ai,\delta_{ai}}^{k_{ai}-1} & 1 & \frac{\bar{\nu}_{ai,1}(s-\lambda_{ai,\delta_{ai}})}{z_{ai,1}-\lambda_{ai,\delta_{ai}}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}(s-\lambda_{ai,\delta_{ai}})}{z_{ai,p_{ai}}-\lambda_{ai,\delta_{ai}}} \end{bmatrix}$$

$$= \frac{1}{\prod_{j=1}^{\delta_{ai}}(s-\lambda_{ai,j})} \det\begin{bmatrix} -\lambda_{ai,1} & \cdots & -\lambda_{ai,1}^{k_{ai}} & 1 & \frac{\bar{\nu}_{ai,1}(s-z_{ai,1})}{z_{ai,1}-\lambda_{ai,1}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}(s-z_{ai,p_{ai}})}{z_{ai,p_{ai}}-\lambda_{ai,1}} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -\lambda_{ai,\delta_{ai}} & \cdots & -\lambda_{ai,\delta_{ai}}^{k_{ai}} & 1 & \frac{\bar{\nu}_{ai,1}(s-z_{ai,1})}{z_{ai,1}-\lambda_{ai,\delta_{ai}}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}(s-z_{ai,p_{ai}})}{z_{ai,p_{ai}}-\lambda_{ai,\delta_{ai}}} \end{bmatrix}$$

$$= \frac{\prod_{j=1}^{p_{ai}}(s-z_{ai,j})}{\prod_{j=1}^{\delta_{ai}}(s-\lambda_{ai,j})} \det\begin{bmatrix} 1 & \lambda_{ai,1} & \cdots & \lambda_{ai,1}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,1}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,1}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{ai,\delta_{ai}} & \cdots & \lambda_{ai,\delta_{ai}}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,\delta_{ai}}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,\delta_{ai}}} \end{bmatrix}$$

Therefore, from (17),

$$\frac{r(s)}{\mu_{ai}(s)} = \frac{\prod_{j=1}^{p_{ai}}(s-z_{ai,j})}{\prod_{j=1}^{k_{ai}+p_{ai}+1}(s-\lambda_{ai,j})} CA^{k_{ai}}F_{ai} \qquad (18)$$

Remark 1. When $\mu_{ai}(t) = \sum_{j=1}^{p_{ai}} \alpha_j e^{z_{ai,j}t}$ where α_j's are arbitrary constants, $r(t)$ becomes zero after a while [11]. Therefore, the fault detection filter cannot detect this type of actuator faults because the residual only has a transient response when the faults occur. For example, if (C, A, F_{ai}) has an invariant zero at the origin, the fault detection filter cannot detect the actuator fault if it is a bias. ◀

3.2 Sensor Fault

From (4) and (6), the transfer matrix from the sensor fault μ_{si} to the residual r is

$$\frac{r(s)}{\mu_{si}(s)} = E_{si} - C(sI-A+LC)^{-1}LE_{si} = C(sI-A+LC)^{-1}\left[(sI-A+LC)f_{si} - LCf_{si}\right]$$

$$= sC(sI-A+LC)^{-1}f_{si} - C(sI-A+LC)^{-1}\bar{f}_{si} \qquad (19)$$

because $E_{si} = Cf_{si}$ and $\bar{f}_{si} = Af_{si}$. Note that $[f_{si}\ \bar{f}_{si}]$ is used for fault detection filter design. When (C, A, f_{si}) has $p_{si,1}$ invariant zeros at $z_{si,1} \cdots z_{si,p_{si,1}}$, the dimension of $\mathcal{T}_{si,1}$, the detection space of f_{si}, is $p_{si,1}+1$ because $Cf_{si} = E_{si} \neq 0$. When $(C, A, \bar{f}_{si})$ has $p_{si,2}$ invariant zeros at $z_{si,p_{si,1}+1} \cdots z_{si,p_{si,1}+p_{si,2}}$, the dimension of $\mathcal{T}_{si,2}$, the detection space of $\bar{f}_{si}$, is $k_{si}+p_{si,2}+1$ where k_{si} is the smallest non-negative integer such that $CA^{k_{si}}\bar{f}_{si} \neq 0$. For the fault detection filter, $\mathcal{T}_{si,1} \oplus \mathcal{T}_{si,2}$ spans $k_{si}+p_{si,1}+p_{si,2}+2$ eigenvectors of $A-LC$. For the fault reconstruction process, it is assumed that $\mathcal{T}_{si,1}$ and $\mathcal{T}_{si,2}$ span $p_{si,1}+1$ and $k_{si}+p_{si,2}+1$ eigenvectors of $A-LC$, respectively. This can be achieved by considering f_{si} and $\bar{f}_{si}$ as two separate faults when designing the fault detection filter. It is also assumed that (C, A, f_{si}) and $(C, A, \bar{f}_{si})$ are mutually detectable.

By following the same derivation in Section 3.1,

$$C(sI-A+LC)^{-1}f_{si} = \frac{\prod_{j=1}^{p_{si,1}}(s-z_{si,j})}{\prod_{j=1}^{p_{si,1}+1}(s-\lambda_{si,j})}Cf_{si}$$

$$C(sI-A+LC)^{-1}\bar{f}_{si} = \frac{\prod_{j=p_{si,1}+1}^{p_{si,1}+p_{si,2}}(s-z_{si,j})}{\prod_{j=p_{si,1}+2}^{k_{si}+p_{si,1}+p_{si,2}+2}(s-\lambda_{si,j})}CA^{k_{si}}\bar{f}_{si}$$

where $\lambda_1 \cdots \lambda_{p_{si,1}+1}$ and $\lambda_{p_{si,1}+2} \cdots \lambda_{k_{si}+p_{si,1}+p_{si,2}+2}$ are the eigenvalues of the fault detection filter associated with $\mathcal{T}_{si,1}$ and $\mathcal{T}_{si,2}$, respectively. Therefore, from (19),

$$\frac{r(s)}{\mu_{si}(s)} = \frac{s\prod_{j=1}^{p_{si,1}}(s-z_{si,j})}{\prod_{j=1}^{p_{si,1}+1}(s-\lambda_{si,j})}E_{si} - \frac{\prod_{j=p_{si,1}+1}^{p_{si,1}+p_{si,2}}(s-z_{si,j})}{\prod_{j=p_{si,1}+2}^{k_{si}+p_{si,1}+p_{si,2}+2}(s-\lambda_{si,j})}CA^{k_{si}}\bar{f}_{si} \tag{20}$$

Remark 2. For certain type of sensor faults, the residual only has a transient response when the faults occur and becomes zero after a while even though the faults still exist. Consider the following system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & 0 \\ A_{21} & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}u$$

$$y = \begin{bmatrix} Cx_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} C & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

For the fault in the sensor that measures x_2, its fault directions are

$$E_s = \begin{bmatrix} 0 \\ 1 \end{bmatrix},\ \begin{bmatrix} f_s & \bar{f}_s \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

Then, the transfer matrix from the sensor fault to the residual is

$$\frac{r(s)}{\mu_s(s)} = \frac{s\prod_{j=1}^{p_{s,1}}(s-z_{s,j})}{\prod_{j=1}^{p_{s,1}+1}(s-\lambda_{s,j})}E_s$$

When the sensor fault is a bias, the residual becomes zero after a while because the transfer matrix has a zero at the origin [11]. Therefore, the fault detection filter cannot detect this type of sensor faults because the residual only has a transient response when the faults occur. Note that the zero at the

origin is not an invariant zero of the fault. One possible example is the bias in a single position sensor, i.e., x_2 is the integral of one of the states x_1. From the physical point of view, this is consistent with the fact that the other states are not affected by the position and only affect the derivative of the position. Therefore, they cannot be used to detect the bias in the position sensor. However, the fault reconstruction process, discussed in Section 4, can still generate the magnitudes of the faults even after the residual becomes zero. This is demonstrated by the numerical example in Section 5.

4 Fault Reconstruction

From (18) and (20), the relationship between the residual and all the actuator and sensor faults can be expressed as

$$r(s) = \sum_{i=1}^{q_a} \left[\frac{\prod_{j=1}^{p_{ai}} (s - z_{ai,j})}{\prod_{j=1}^{k_{ai}+p_{ai}+1} (s - \lambda_{ai,j})} C A^{k_{ai}} F_{ai} \right] \mu_{ai}(s) + \sum_{i=1}^{q_s} \left[\frac{s \prod_{j=1}^{p_{si,1}} (s - z_{si,j})}{\prod_{j=1}^{p_{si,1}+1} (s - \lambda_{si,j})} E_{si} - \frac{\prod_{j=p_{si,1}+1}^{p_{si,1}+p_{si,2}} (s - z_{si,j})}{\prod_{j=p_{si,1}+2}^{k_{si}+p_{si,1}+p_{si,2}+2} (s - \lambda_{si,j})} C A^{k_{si}} \bar{f}_{si} \right] \mu_{si}(s) \tag{21}$$

In Section 4.1, the reconstruction of the actuator fault is discussed. In Section 4.2, the reconstruction of the sensor fault is discussed.

4.1 Actuator Fault

In order to reconstruct the actuator fault μ_{ai}, a projected residual that is only sensitive to μ_{ai}, but not to the other faults, is needed. Define a projector $\hat{H}_{ai}$ that annihilates all the faults except μ_{ai}.

$$\hat{H}_{ai} = I - \operatorname{Ker} \hat{H}_{ai} \left[(\operatorname{Ker} \hat{H}_{ai})^T \operatorname{Ker} \hat{H}_{ai} \right]^{-1} (\operatorname{Ker} \hat{H}_{ai})^T \tag{22}$$

where $\operatorname{Ker} \hat{H}_{ai} = \operatorname{Im}[\, C A^{k_{a1}} F_{a1} \;\; C A^{k_{a2}} F_{a2} \;\cdots\; C A^{k_{a,i-1}} F_{a,i-1} \;\; C A^{k_{a,i+1}} F_{a,i+1} \;\cdots\; C A^{k_{a,q_a}} F_{a,q_a} \;\; E_{s1} \;\; E_{s2} \;\cdots\; E_{s,q_s} \;\; C A^{k_{s1}} \bar{f}_{s1} \;\; C A^{k_{s2}} \bar{f}_{s2} \;\cdots\; C A^{k_{s,q_s}} \bar{f}_{s,q_s} \,]$. Note that $\hat{H}_{ai}$ is the same as the projector (10) used by the fault detection filter. By operating $\hat{H}_{ai}$ on the residual, (21) becomes

$$\hat{H}_{ai} r(s) = \frac{\prod_{j=1}^{p_{ai}} (s - z_{ai,j})}{\prod_{j=1}^{k_{ai}+p_{ai}+1} (s - \lambda_{ai,j})} \hat{H}_{ai} C A^{k_{ai}} F_{ai} \mu_{ai}(s)$$

Therefore, the projected residual $\hat{H}_{ai} r$ is only sensitive to μ_{ai}, but not to the other faults. Let q_{ai} be a m by 1 vector where m is the number of the measurements. By operating q_{ai} on $\hat{H}_{ai} r$, the actuator fault μ_{ai} can be reconstructed from the projected residual $\hat{H}_{ai} r$ by using

$$\mu_{ai}(s) = \frac{1}{q_{ai}^T \hat{H}_{ai} C A^{k_{ai}} F_{ai}} \frac{\prod_{j=1}^{k_{ai}+p_{ai}+1} (s - \lambda_{ai,j})}{\prod_{j=1}^{p_{ai}} (s - z_{ai,j})} q_{ai}^T \hat{H}_{ai} r(s) \tag{23}$$

if all the invariant zeros of (C, A, F_{ai}) are in the left-half plane. Since $CA^{k_{ai}}F_{ai} \not\subseteq \text{Ker}\,\hat{H}_{ai}$, $\hat{H}_{ai}CA^{k_{ai}}F_{ai} \neq 0$ and there exists q_{ai} such that $q_{ai}^T\hat{H}_{ai}CA^{k_{ai}}F_{ai} \neq 0$. For example, $q_{ai} = \hat{H}_{ai}CA^{k_{ai}}F_{ai}$. Note that the Silverman's algorithm requires the left inverse of $\hat{H}_{ai}CA^{k_{ai}}F_{ai}$ [8]. Also note that (23) is not proper. In order to avoid differentiating and amplifying the disturbance, a $(k_{ai}+1)$-dimensional low-pass filter with poles assigned as butterworth configuration may be used at the expense of introducing a delay in reconstructing the actuator fault.

Since q_{ai} is not unique, an optimization problem is formulated to determine q_{ai} by considering the disturbance. Consider the system (4) with the colored noise w,

$$\dot{x} = Ax + Bu + B_w w + \sum_{i=1}^{q_a} F_{ai}\mu_{ai}$$

$$y = Cx + D_w w + \sum_{i=1}^{q_s} E_{si}\mu_{si}$$

where $\dot{w} = \bar{A}w + \bar{w}$ and $\bar{w}$ is the white noise. Then,

$$q_{ai}^T\hat{H}_{ai}r(s) = q_{ai}^T G_{\mu_{ai}}(s)\mu_{ai}(s) + q_{ai}^T G_{w_{ai}}(s)\bar{w}(s)$$

where $G_{\mu_{ai}}(s) = \frac{\prod_{j=1}^{p_{ai}}(s-z_{ai,j})}{\prod_{j=1}^{k_{ai}+p_{ai}+1}(s-\lambda_{ai,j})}\hat{H}_{ai}CA^{k_{ai}}F_{ai}$ and $G_{w_{ai}}(s) = \hat{H}_{ai}[C(sI-A+LC)^{-1}(B_w - LD_w) + D_w]$ $(sI-\bar{A})^{-1}$. The optimal q_{ai} is determined by minimizing the ratio of the $\mathcal{H}_2$ norm of the transfer matrix from $\bar{w}$ to $q_{ai}^T\hat{H}_{ai}r$ over the $\mathcal{H}_2$ norm of the transfer function from μ_{ai} to $q_{ai}^T\hat{H}_{ai}r$.

$$\min_{q_{ai}} J = \min_{q_{ai}} \frac{\|q_{ai}^T G_{w_{ai}}\|_2^2}{\|q_{ai}^T G_{\mu_{ai}}\|_2^2} \tag{24}$$

This minimization problem can be solved by rewriting the cost criterion as

$$\bar{J} = \|q_{ai}^T G_{w_{ai}}\|_2^2 - \gamma\|q_{ai}^T G_{\mu_{ai}}\|_2^2 = q_{ai}^T\bar{G}_{w_{ai}}q_{ai} - \gamma q_{ai}^T\bar{G}_{\mu_{ai}}q_{ai}$$

where γ is a Lagrange multiplier, $\bar{G}_{w_{ai}} = \frac{1}{2\pi}\int_{-\infty}^{\infty} G_{w_{ai}}(jw)G_{w_{ai}}^T(-jw)\,dw$ and $\bar{G}_{\mu_{ai}} = \frac{1}{2\pi}\int_{-\infty}^{\infty} G_{\mu_{ai}}(jw)$ $G_{\mu_{ai}}^T(-jw)\,dw$. Note that $\bar{G}_{w_{ai}}$ and $\bar{G}_{\mu_{ai}}$ can be computed by using their state-space models [12]. For example, $\bar{G}_{\mu_{ai}} = \hat{H}_{ai}CW_{\mu_{ai}}C^T\hat{H}_{ai}$ where $W_{\mu_{ai}}$ is the controllability gramian of $(A-LC, F_{ai})$ and $\bar{G}_{w_{ai}} = \hat{H}_{ai}$ $[C\ \ D_w]W_{w_{ai}}[C\ \ D_w]^T\hat{H}_{ai}$ where $W_{w_{ai}}$ is the controllability gramian of $\left(\begin{bmatrix} A-LC & B_w-LD_w \\ 0 & \bar{A} \end{bmatrix}, \begin{bmatrix} 0 \\ I \end{bmatrix}\right)$. From the first-order necessary condition,

$$\frac{\partial\bar{J}}{\partial q_{ai}} = q_{ai}^T\bar{G}_{w_{ai}} - \gamma q_{ai}^T\bar{G}_{\mu_{ai}} = 0 \Rightarrow \bar{G}_{w_{ai}}q_{ai} = \gamma\bar{G}_{\mu_{ai}}q_{ai} \tag{25}$$

Therefore, the optimal q_{ai} is the generalized eigenvector of $(\bar{G}_{w_{ai}}, \bar{G}_{\mu_{ai}})$ associated with the smallest generalized eigenvalue and the optimal J is the smallest generalized eigenvalue. Note that the ranks of $\bar{G}_{w_{ai}}$ and $\bar{G}_{\mu_{ai}}$ are $m - \dim(\text{Ker}\,\hat{H}_{ai})$ and 1, respectively. To solve for the generalized eigenvector, it is more numerically robust if the dimension of $\bar{G}_{w_{ai}}$ and $\bar{G}_{\mu_{ai}}$ is reduced from m to $m - \dim(\text{Ker}\,\hat{H}_{ai})$.

4.2 Sensor Fault

In order to obtain a projected residual that is only sensitive to the sensor fault μ_{si}, but not to the other faults, a projector $\hat{H}_{si,1}$ is defined as

$$\hat{H}_{si,1} = I - \mathrm{Ker}\hat{H}_{si,1}\left[(\mathrm{Ker}\hat{H}_{si,1})^T\mathrm{Ker}\hat{H}_{si,1}\right]^{-1}(\mathrm{Ker}\hat{H}_{si,1})^T \tag{26}$$

where $\mathrm{Ker}\,\hat{H}_{si,1} = \mathrm{Im}\,[CA^{k_{a1}}F_{a1}\ CA^{k_{a2}}F_{a2}\cdots CA^{k_{a,q_a}}F_{a,q_a}\ E_{s1}\ E_{s2}\cdots E_{s,q_s}\ CA^{k_{s1}}\bar{f}_{s1}\ CA^{k_{s2}}\bar{f}_{s2}\cdots CA^{k_{s,i-1}}\bar{f}_{s,i-1}\ CA^{k_{s,i+1}}\bar{f}_{s,i+1}\cdots CA^{k_{s,q_s}}\bar{f}_{s,q_s}]$. Note that $\hat{H}_{si,1}$ is different from the projector (10) used by the fault detection filter where E_{si} is not in the null space of the projector and now it is. By following the same derivation in Section 4.1, the sensor fault μ_{si} can be reconstructed from the projected residual $\hat{H}_{si,1}r$ by using

$$\mu_{si}(s) = \frac{1}{q_{si,1}^T\hat{H}_{si,1}CA^{k_{si}}\bar{f}_{si}}\frac{\prod_{j=p_{si,1}+2}^{k_{si}+p_{si,1}+p_{si,2}+2}(s-\lambda_{si,j})}{\prod_{j=p_{si,1}+1}^{p_{si,1}+p_{si,2}}(s-z_{si,j})}q_{si,1}^T\hat{H}_{si,1}r(s) \tag{27}$$

if all the invariant zeros of $(C, A, \bar{f}_{si})$ are in the left-half plane. The optimal $q_{si,1}$ can be determined similarly as in Section 4.1. Note that (27) is not proper and a $(k_{si}+1)$-dimensional low-pass filter may be used to reduce the effect of the disturbance at the expense of introducing a delay in reconstructing the sensor fault.

There is an alternative approach to reconstruct the sensor fault μ_{si}. Define a projector $\hat{H}_{si,2}$ as

$$\hat{H}_{si,2} = I - \mathrm{Ker}\hat{H}_{si,2}\left[(\mathrm{Ker}\hat{H}_{si,2})^T\mathrm{Ker}\hat{H}_{si,2}\right]^{-1}(\mathrm{Ker}\hat{H}_{si,2})^T \tag{28}$$

where $\mathrm{Ker}\,\hat{H}_{si,2} = \mathrm{Im}\,[CA^{k_{a1}}F_{a1}\ CA^{k_{a2}}F_{a2}\cdots CA^{k_{a,q_a}}F_{a,q_a}\ E_{s1}\ E_{s2}\cdots E_{s,i-1}\ E_{s,i+1}\cdots E_{s,q_s}\ CA^{k_{s1}}\bar{f}_{s1}\ CA^{k_{s2}}\bar{f}_{s2}\cdots CA^{k_{s,q_s}}\bar{f}_{s,q_s}]$. Then, the sensor fault μ_{si} can also be reconstructed from the projected residual $\hat{H}_{si,2}r$ by using

$$\mu_{si}(s) = \frac{1}{q_{si,2}^T\hat{H}_{si,2}E_{si}}\frac{\prod_{j=1}^{p_{si,1}+1}(s-\lambda_{si,j})}{s\prod_{j=1}^{p_{si,1}}(s-z_{si,j})}q_{si,2}^T\hat{H}_{si,2}r(s) \tag{29}$$

if all the invariant zeros of (C, A, f_{si}) are in the left-half plane. The optimal $q_{si,2}$ can be determined similarly as in Section 4.1 except a finite frequency range of the designer's choice would be used instead of the frequency range from $-\infty$ to ∞ because the transfer matrix from μ_{si} to $q_{si,2}^T\hat{H}_{si,2}r$ is not strictly proper. Note that (29) is proper. The reconstructed sensor fault generated by (29) may be less sensitive to the disturbance than the one generated by (27) because the disturbance is not differentiated. Furthermore, since a low-pass filter is not required for (29), there is no delay in reconstructing the sensor fault. However, (29) is only stable in the sense of Lyapunov, but not asymptotically stable. The effect of the disturbance may accumulate over time.

Remark 3. Consider a linear time-invariant system that is observable with all sensors, but unobservable without one of the sensors.

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u$$
$$y = \begin{bmatrix} Cx_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} C & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

This system is unobservable without the sensor that measures x_2. For the fault in this sensor, its fault directions are

$$E_s = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} f_s & \bar{f}_s \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & A_{22} \end{bmatrix}$$

Then, the transfer matrix from the sensor fault to the residual is

$$\frac{r(s)}{\mu_s(s)} = \frac{(s - A_{22}) \prod_{j=1}^{p_{s,1}} (s - z_{s,j})}{\prod_{j=1}^{p_{s,1}+1} (s - \lambda_{s,j})} E_s$$

Hence, the eigenvalue associated with the unobservable mode will become one of the poles of the fault reconstruction process. Therefore, for reconstructing a sensor fault, the system has to be detectable with respect to the other sensors. ◀

5 Numerical Example

Consider a linear time-invariant system with

$$A = \begin{bmatrix} -4 & 1 & 2 & -3 & 5 & 0 \\ 2 & -3 & -5 & 0 & 2 & 0 \\ -3 & 2 & -7 & 1 & -4 & 0 \\ 5 & -1 & 3 & -2 & 5 & 0 \\ -3 & -5 & 1 & -4 & -8 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 2 \\ 1 \\ -5 \\ 3 \\ -4 \\ 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

A fault detection filter is designed to detect and identify the faults in the actuator, second sensor and fifth sensor. Note that the fifth sensor can be considered as a position sensor because it measures the sixth state which is the integral of the fifth state and does not affect other states. From (4) and (5), the fault directions

are

$$F_a = \begin{bmatrix} 2 \\ 1 \\ -5 \\ 3 \\ -4 \\ 0 \end{bmatrix}, \quad E_{s1} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad E_{s2} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} f_{s1} & \bar{f}_{s1} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & -3 \\ 0 & 2 \\ 0 & -1 \\ 0 & -5 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} f_{s2} & \bar{f}_{s2} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}$$

By using the design algorithm in [5], the fault detection filter gain is obtained as

$$L = \begin{bmatrix} -21.4426 & 1.0000 & 13.2130 & 23.9835 & 0.0000 \\ -6.9663 & -0.5000 & -0.1685 & 12.1966 & 0.0000 \\ 11.2900 & 2.0000 & -9.4501 & -15.6101 & -0.0000 \\ -14.8214 & -1.0000 & 13.1070 & 26.3927 & 0.0000 \\ 16.2212 & -5.0000 & -8.1061 & -27.9910 & -0.0000 \\ -4.3015 & 0.0000 & 2.5077 & 5.7139 & 3.5000 \end{bmatrix}$$

The eigenvalue associated with F_a is -5. The eigenvalues associated with f_{s1} and $\bar{f}_{s1}$ are -2.5 and -3, respectively. The eigenvalue associated with f_{s2} is -3.5.

To evaluate the performance of the fault detection filter, an actuator fault and two sensor faults are imposed on the system separately. The actuator fault simulates a stuck actuator. In Figure 2, the top left figure shows the control command. The middle left figure shows the actuator fault μ_a which occurs at sixth second. The bottom left figure shows the control input applied to the system, which is the sum of the control command and the actuator fault. It shows that the actuator is stuck at 1 after sixth second regardless of the control command. The sensor faults simulate the bias developed in the sensors. In Figures 3 and 4, the top left figures show the second sensor fault μ_{s1} and the fifth sensor fault μ_{s2} which start at the fourth second and end at the twelfth second, respectively. Figure 5 shows the time response of the norms of the three projected residuals generated by the fault detection filter (6) using projectors (10) in the presence of the colored sensor noise where $\bar{A} = -1000I$, the power spectral density of $\bar{w}$ is $2I$, $B_w = 0$ and $D_w = I$. Each row shows the projected residuals when one of the faults occurs. Each column shows one of the projected residuals when the faults occur. Note that only the projected residual associated with the faulty instrument becomes large when the fault occurs. However, the projected residual associated the fifth sensor becomes small after a while even though the fifth sensor fault still exists. This is consistent with the discussion in Remark 2. Therefore, the fault detection filter can detect and identify the actuator and second sensor faults, but not the fifth sensor fault.

To reconstruct these three faults, the relationship between the residual and the faults is obtained from (21).

$$r(s) = \frac{1}{s+5} CF_a \mu_a(s) + \left(\frac{s}{s+2.5} E_{s1} - \frac{1}{s+3} C\bar{f}_{s1} \right) \mu_{s1}(s) + \frac{s}{s+3.5} E_{s2} \mu_{s2}(s)$$

From (22), the projector $\hat{H}_a$ used for reconstructing the actuator fault is obtained by annihilating $[E_{s1} \;\; C\bar{f}_{s1} \;\; E_{s2}]$. From (23), the actuator fault can be reconstructed from

$$\mu_a(s) = \frac{s+5}{q_a^T \hat{H}_a C F_a} q_a^T \hat{H}_a r(s) \tag{30}$$

with the optimal $q_a = [0.9114 \;\; 0 \;\; -0.3903 \;\; 0.1308 \;\; 0]$ determined from (25). In Figure 2, the top middle figure shows the reconstructed actuator fault generated by (30). To reduce the effect of the sensor noise, a low-pass filter is added to (30).

$$\hat{\mu}_a(s) = \frac{20(s+5)}{q_a^T \hat{H}_a C F_a (s+20)} q_a^T \hat{H}_a r(s) \tag{31}$$

The initial condition of the fault reconstruction process is zero given that there is no fault initially. The middle center figure shows the reconstructed actuator fault generated by (31) which is close to the actual actuator fault shown in the middle left figure. By adding this reconstructed actuator fault to the control command, the control input applied to the system is reconstructed and shown in the bottom middle figure which is close to the actual control input shown in the bottom left figure. This information can be used to evaluate the condition of the actuator and in this case, the actuator found to be stuck. To demonstrate that the reconstructed actuator fault generated with an arbitrarily chosen q_a is more sensitive to the sensor noise than the one generated with the q_a derived from solving (24), the top right figure shows the reconstructed actuator fault generated by (30) with q_a arbitrarily chosen as $[0 \;\; 0 \;\; 0 \;\; 1 \;\; 0]$. To reduce the effect of the sensor noise, a low-pass filter with a slower pole is used.

$$\hat{\mu}_a(s) = \frac{6(s+5)}{q_a^T \hat{H}_a C F_a (s+6)} q_a^T \hat{H}_a r(s) \tag{32}$$

The middle right figure shows the reconstructed actuator fault generated by (32) whose delay in reconstructing the actuator fault is worse than (31). This becomes clearer in the bottom right figure where the reconstructed control input is shown.

From (26) and (28), two projectors used for reconstructing the second sensor fault are obtained by annihilating $[CF_a \;\; E_{s1} \;\; E_{s2}]$ and $[CF_a \;\; C\bar{f}_{s1} \;\; E_{s2}]$, respectively. In Figure 3, the middle left and bottom left figures show the reconstructed second sensor faults generated by (27) with a low-pass filter and (29), respectively. Both are close to the actual sensor fault shown in the top left figure. However, the reconstructed sensor fault generated by (27) is more sensitive to the sensor noise and has a delay due to the low-pass filter. By subtracting the reconstructed sensor faults from the faulty measurement, the second measurements are

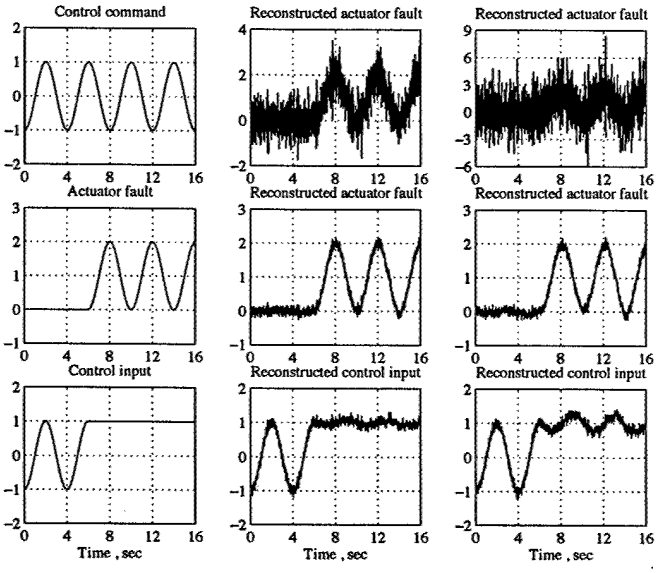


Figure 2: Fault reconstruction for the actuator

reconstructed and shown in the middle and bottom right figures which are close to the correct measurement shown in the top right figure. In the middle right figure, the spikes at fourth and twelfth second are due to the delay in reconstructing the sensor fault.

For reconstructing the fifth sensor fault, the projector can only be obtained from (28) by annihilating $[CF_a \;\; E_{s1} \;\; C\bar{f}_{s1}]$. In Figure 4, the bottom left figure shows the reconstructed sensor fault generated by (29) which is close to the actual sensor fault shown in the top left figure. The bottom right figure shows the reconstructed fifth measurement which is close to the correct measurement shown in the top right figure. Note that the fault reconstruction process can still generate the magnitude of the fifth sensor fault even after the projected residual becomes zero as shown in the bottom right figure of Figure 5.

6 Conclusion

The fault reconstruction process generates the magnitudes of sensor and actuator faults using the residual generated by the fault detection filter. An optimal fault reconstruction process is derived from solving a minimization problem by considering the disturbance. For the existence of the fault reconstruction process, the invariant zeros of the fault have to be in the left-half plane. Furthermore, for reconstructing a sensor fault, the system has to be detectable with respect to the other sensors. Although the fault reconstruction process can also be derived numerically by using the Silverman's algorithm, it is not optimal in general and its existence conditions and analytical structure cannot be obtained.

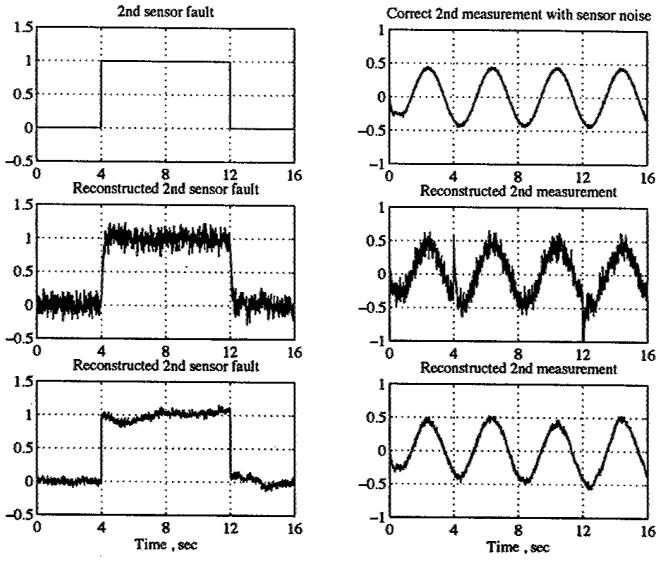


Figure 3: Fault reconstruction for the 2nd sensor

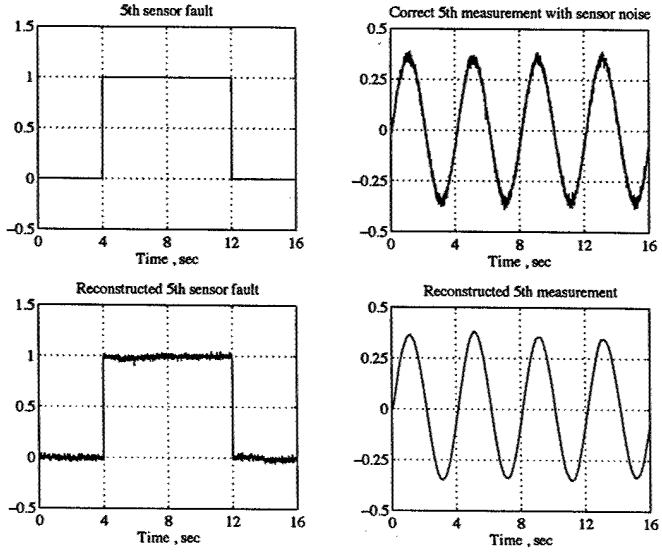


Figure 4: Fault reconstruction for the 5th sensor

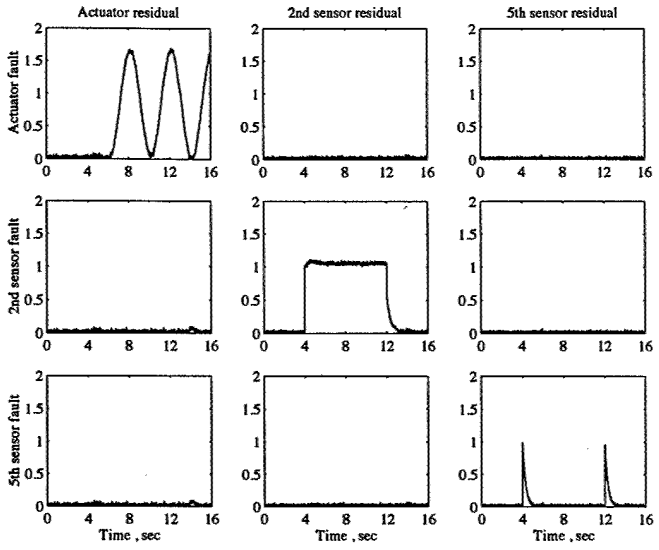


Figure 5: Projected residuals generated by the fault detection filter

Appendix

If the first eigenvalue of $A - LC$ associated with $\mathcal{T}_{ai}$ is at the first invariant zero of (C, A, F_{ai}), i.e., $\lambda_{ai,1} = z_{ai,1}$, (14) is still true except when $j = k = 1$, (14) becomes $\bar{\nu}_{ai,1}\alpha_{1,1} = 0$. If $\bar{\nu}_{ai,1} = 0$, (15) and (16) become

$$\begin{bmatrix} F_{ai} & AF_{ai} & \cdots & A^{k_{ai}}F_{ai} & \nu_{ai,1} & \cdots & \nu_{ai,p_{ai}} \end{bmatrix} = \begin{bmatrix} x_1 & \alpha_{2,1}x_2 & \cdots & \alpha_{\delta_{ai},1}x_{\delta_{ai}} \end{bmatrix}$$
$$\begin{bmatrix} \alpha_{1,1} & \alpha_{1,1}\lambda_{ai,1} & \cdots & \alpha_{1,1}\lambda_{ai,1}^{k_{ai}} & \alpha_{1,k_{ai}+2} & \frac{\alpha_{1,1}\bar{\nu}_{ai,2}}{z_{ai,2}-\lambda_{ai,1}} & \cdots & \frac{\alpha_{1,1}\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,1}} \\ 1 & \lambda_{ai,2} & \cdots & \lambda_{ai,2}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,2}} & \frac{\bar{\nu}_{ai,2}}{z_{ai,2}-\lambda_{ai,2}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,2}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{ai,\delta_{ai}} & \cdots & \lambda_{ai,\delta_{ai}}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,\delta_{ai}}} & \frac{\bar{\nu}_{ai,2}}{z_{ai,2}-\lambda_{ai,\delta_{ai}}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,\delta_{ai}}} \end{bmatrix}$$

and

$$C(sI - A + LC)^{-1}F_{ai} = C\begin{bmatrix} x_1 & \alpha_{2,1}x_2 & \cdots & \alpha_{\delta_{ai},1}x_{\delta_{ai}} \end{bmatrix}\begin{bmatrix} \frac{\alpha_{1,1}}{s-\lambda_{ai,1}} & \frac{1}{s-\lambda_{ai,2}} & \cdots & \frac{1}{s-\lambda_{ai,\delta_{ai}}} \end{bmatrix}^T$$

Then, by following the same derivation in Section 3.1, (18) can be derived with a pole-zero cancellation. If $\alpha_{1,1} = 0$, (15) and (16) become

$$\begin{bmatrix} F_{ai} & AF_{ai} & \cdots & A^{k_{ai}}F_{ai} & \nu_{ai,1} & \cdots & \nu_{ai,p_{ai}} \end{bmatrix} = \begin{bmatrix} \alpha_{1,k_{ai}+2}x_1 & \alpha_{2,1}x_2 & \cdots & \alpha_{\delta_{ai},1}x_{\delta_{ai}} \end{bmatrix}$$
$$\begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 1 & \lambda_{ai,2} & \cdots & \lambda_{ai,2}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,2}} & \frac{\bar{\nu}_{ai,2}}{z_{ai,2}-\lambda_{ai,2}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,2}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{ai,\delta_{ai}} & \cdots & \lambda_{ai,\delta_{ai}}^{k_{ai}} & \frac{\bar{\nu}_{ai,1}}{z_{ai,1}-\lambda_{ai,\delta_{ai}}} & \frac{\bar{\nu}_{ai,2}}{z_{ai,2}-\lambda_{ai,\delta_{ai}}} & \cdots & \frac{\bar{\nu}_{ai,p_{ai}}}{z_{ai,p_{ai}}-\lambda_{ai,\delta_{ai}}} \end{bmatrix}$$

and

$$C(sI - A + LC)^{-1}F_{ai} = C \left[\begin{array}{cccc} \alpha_{1,k_{ai}+2}x_1 & \alpha_{2,1}x_2 & \cdots & \alpha_{\delta_{ai},1}x_{\delta_{ai}} \end{array} \right] \left[\begin{array}{cccc} 0 & \frac{1}{s-\lambda_{ai,2}} & \cdots & \frac{1}{s-\lambda_{ai,\delta_{ai}}} \end{array} \right]^T$$

Then, by following the same derivation in Section 3.1, (18) can also be derived with a pole-zero cancellation. If $\bar{\nu}_{ai,1} = \alpha_{1,1} = 0$, the derivation is similar. Note that the two derivations above can be extended to the case where multiple eigenvalues of $A - LC$ associated with $\mathcal{T}_{ai}$ are at the invariant zeros of (C, A, F_{ai}).

References

[1] R. V. Beard, *Failure Accomodation in Linear Systems through Self-Reorganization*, Ph.D. thesis, Massachusetts Institute of Technology, 1971.

[2] H. L. Jones, *Failure Detection in Linear Systems*, Ph.D. thesis, Massachusetts Institute of Technology, 1973.

[3] Mohammad-Ali Massoumnia, "A geometric approach to the synthesis of failure detection filters," *IEEE Transactions on Automatic Control*, vol. AC-31, no. 9, pp. 839–846, Sept. 1986.

[4] John E. White and Jason L. Speyer, "Detection filter design: Spectral theory and algorithms," *IEEE Transactions on Automatic Control*, vol. AC-32, no. 7, pp. 593–603, July 1987.

[5] Randal K. Douglas and Jason L. Speyer, "Robust fault detection filter design," *AIAA Journal of Guidance, Control, and Dynamics*, vol. 19, no. 1, pp. 214–218, Jan-Feb 1996.

[6] Randal K. Douglas and Jason L. Speyer, "$\mathcal{H}_\infty$ bounded fault detection filter," *AIAA Journal of Guidance, Control, and Dynamics*, vol. 22, no. 1, pp. 129–138, Jan-Feb 1999.

[7] Leonard M. Silverman, "Inversion of multivariable linear systems," *IEEE Transactions on Automatic Control*, vol. AC-14, no. 3, pp. 270–276, June 1969.

[8] L. M. Silverman and H. J. Payne, "Input-output structure of linear systems with application to the decoupling problem," *SIAM Journal of Control*, vol. 9, no. 2, pp. 199–233, May 1971.

[9] Walter H. Chung and Jason L. Speyer, "A game theoretic fault detection filter," *IEEE Transactions on Automatic Control*, vol. AC-43, no. 2, pp. 143–161, Feb. 1998.

[10] Thomas Kailath, *Linear Systems*, Prentice Hall, 1980.

[11] Chi-Tsong Chen, *Linear system theory and design*, Holt, Rinehart, and Winston, 1984.

[12] John C. Doyle, Keith Glover, Pramod P. Khargonekar, and Bruce A. Francis, "State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ Control Problems," *IEEE Transactions on Automatic Control*, vol. AC-34, no. 8, pp. 831–847, Aug. 1989.

Appendix F

"Robust Multiple-Fault Detection Filter,"

Robert H. Chen and Jason L. Speyer,

The special issue of condition monitoring, fault detection and isolation in the *International Journal of Robust and Nonlinear Control*, Vol. 12, Issue 8, 2002.

INTERNATIONAL JOURNAL OF ROBUST AND NONLINEAR CONTROL
Int. J. Robust Nonlinear Control 2002; **12**:675–696 (DOI: 10.1002/rnc.715)

Robust multiple-fault detection filter

Robert H. Chen and Jason L. Speyer*,†

Mechanical and Aerospace Engineering Department, University of California, Los Angeles, CA 90095-1597, USA

SUMMARY

A new robust multiple-fault detection and identification algorithm is determined. Different from other algorithms which explicitly force the geometric structure by using eigenstructure assignment or geometric theory, this algorithm is derived from solving an optimization problem. The output error is divided into several subspaces. For each subspace, the transmission from one fault, denoted the associated target fault, is maximized while the transmission from other faults, denoted the associated nuisance fault, is minimized. Therefore, each projected residual of the robust multiple-fault detection filter is affected primarily by one fault and minimally by other faults. The transmission from process and sensor noises is also minimized so that the filter is robust with respect to these disturbances. It is shown that, in the limit where the weighting on each associated nuisance fault transmission goes to infinity, the filter recovers the geometric structure of the restricted diagonal detection filter of which the Beard–Jones detection filter and unknown input observer are special cases. Filter designs can be obtained for both time-invariant and time-varying systems.

KEY WORDS: fault detection and identification; analytical redundancy; Beard–Jones detection filter; approximate fault detection filter; robust fault detection filter; time-varying system

1. INTRODUCTION

Any system under automatic control demands a high degree of reliability in order to operate properly. This requires a health monitoring system capable of detecting any plant, actuator and sensor faults as they occur and identifying the faulty components. This process is called fault detection and identification. The most common approach to fault detection and identification is hardware redundancy which is the direct comparison of the outputs from identical components. It requires very little computation. However, hardware redundancy is expensive and limited by space and weight. An alternative is analytical redundancy which uses

*Correspondence to: Jason L. Speyer, Mechanical and Aerospace Engineering Department, University of California, Los Angeles, Los Angeles, California 90095-1597, USA.
†E-mail: speyer@seas.ucla.edu

Contract/grant sponsor: Air Force Office of Scientific Research; contact/grant number: F49620-00-1-0154.
Contract/grant sponsor: NASA Goddard Space Flight Center; contract/grant number: NAG5-11384.
Contract/grant sponsor: California Department of Transportation; contract/grant number: TO 4209

the modelled dynamic relationship between system inputs and measured system outputs to form a residual process. Nominally, the residual is non-zero only when a fault has occurred and is zero at other times. Therefore, no redundant components are needed. However, additional computation is required.

A popular approach to analytical redundancy is the detection filter which was first introduced by Beard [1] and refined by Jones [2]. It is also known as Beard–Jones detection (BJD) filter. A geometric interpretation and a spectral analysis of the BJD filter are given in References [3,4], respectively. The idea of the BJD filter is to place the reachable subspace of each fault into invariant subspaces which do not overlap each other. Then, when a non-zero residual is detected, a fault can be announced and identified by projecting the residual onto each of the invariant subspaces. In this way, multiple faults can be monitored in one filter. A design algorithm [5] improves the robustness of the BJD filter by imposing the geometric structure to isolate the faults and using the design freedom remaining to bound the process and sensor noise transmission.

In Reference [3], a more general form of the detection filter, called restricted diagonal detection (RDD) filter, is given of which the BJD filter is a special case. Instead of placing each fault into an invariant subspace like the BJD filter does, the RDD filter places all the other faults associated with each fault that needs to be detected into the unobservable subspace of a projected residual. Therefore, each projected residual is only sensitive to one fault, but not to the other faults. When every fault is detected, the RDD filter is equivalent to the BJD filter. However, some faults do not need to be detected, but only need to be blocked from the projected residuals. For example, certain process noise and plant certainty may be modelled as faults. By relaxing the constraint on detecting the faults that do not need to be detected, the RDD filter is more robust than the BJD filter [6].

One related approach, unknown input observer [7–9], is another special case of the RDD filter when only one fault is detected. The faults are divided into two groups: a single target fault and possibly several nuisance faults. The nuisance faults are placed in the unobservable subspace of the residual. Therefore, the residual is only sensitive to the target fault, but not to the nuisance faults. Although only one fault can be monitored in each unknown input observer, there are some benefits. For example, one gains additional flexibility which can be used to improve robustness and time-varying systems can be treated [10–12].

In this paper, a new robust multiple-fault detection and identification algorithm is derived from solving an optimization problem. The output error is divided into several subspaces by using projectors. For each subspace, the projected output error variance due to one fault, denoted the associated target fault, is maximized and the projected output error variance due to other faults, denoted the associated nuisance fault, process noise, sensor noise and initial conditional error is minimized. The cost criterion is constructed such that each projected output error variance is included as a sum which produces approximately the geometric structure of the RDD filter. Therefore, each projected residual of the robust multiple-fault detection filter is affected primarily by one fault and minimally by other faults and is robust with respect to the disturbances. Note that [12], an approximate unknown input observer, is a special case of the filter when only one fault is detected.

In the limit where the weighting on each projected output error variance due to the associated nuisance fault goes to infinity, it is shown that the filter places each associated nuisance fault into the unobservable subspace of its associated projected residual when there is no

complementary subspace[‡] for both time-invariant and time-varying systems. Therefore, the filter becomes equivalent to the RDD filter in the limit and extends the RDD filter to the time-varying case. Numerical examples show that the filter is an approximate RDD filter when it is not in the limit even if there exists the complementary subspace. These limiting results are important in ensuring that both fault detection and identification can occur.

The robust multiple-fault detection filter is fundamentally different from other design algorithms for the RDD or BJD filter which explicitly force the geometric structure by using eigenstructure assignment [4,6] or geometric theory [3,5]. Rather, the filter is derived from solving an optimization problem and only in the limit, is the geometric structure of the RDD filter recovered and the faults are completely isolated. When it is not in the limit, the filter only isolates the faults within approximate unobservable subspaces. This new feature allows the filter to be potentially more robust because of the additional design freedom which allows different degrees of fault isolation. Furthermore, a mechanism that enhances the sensitivity of the projected residuals to their associated target faults is provided. Finally, the filter can be applied to time-varying systems. Although the filter has all these advantages, the process of deriving the filter gain requires the solution to a two-point boundary value problem which includes a set of Lyapunov equations. However, the filter gain computation can be done off-line so that the filter implementation is as straightforward as the RDD filter.

The problem is formulated in Section 2 and its solution is derived in Section 3. In Section 4, the filter is determined in the limit when there is no complementary subspace. In Section 5, the projectors used to divide the output error are derived from solving the optimization problem. In Section 6, numerical examples are given.

2. PROBLEM FORMULATION

Consider a linear, time-varying, uniformly observable system

$$\dot{x} = Ax + B_u u + B_w w \tag{1a}$$

$$y = Cx + v \tag{1b}$$

where u is the control input, y is the measurement, w is the process noise and v is the sensor noise. Following the development in References [1,4,10], any plant, actuator and sensor faults can be modelled as additive terms in the state equation (1a). Therefore, a linear system with q faults can be modelled by

$$\dot{x} = Ax + B_u u + B_w w + \sum_{i=1}^{q} F_i \mu_i \tag{2a}$$

$$y = Cx + v \tag{2b}$$

The fault magnitudes μ_i are unknown and arbitrary functions of time that are zero when there is no fault. The fault directions F_i are maps that are *a priori* known. Assume F_i are monic so that $\mu_i \neq 0$ imply $F_i \mu_i \neq 0$.

[‡]The union of the invariant subspace of each fault fills the entire state space leaving no remaining subspace, the complementary subspace.

If the first s faults need to be detected where $s \leqslant q$, the objective of the robust multiple-fault detection filter problem is to find a filter gain L for the linear observer

$$\dot{\hat{x}} = A\hat{x} + B_u u + L(y - C\hat{x}) \tag{3}$$

and projectors $\hat{H}_1 \cdots \hat{H}_s$ which operate on the residual

$$r = y - C\hat{x} \tag{4}$$

such that each projected residual $\hat{H}_i r$ is affected primarily by its associated target fault μ_i and minimally by its associated nuisance fault $\hat{\mu}_i = [\mu_1 \cdots \mu_{i-1}\ \mu_{i+1} \cdots \mu_q]^T$, process noise w, sensor noise v and initial condition error $x(t_0) - \hat{x}(t_0)$. This approximates the RDD filter problem. Even though the last $q - s$ faults are not detected, they are blocked from the s projected residuals used for detecting the first s faults. By relaxing the constraint on detecting the last $q - s$ faults, the robustness of the filter is improved [6]. When $s = q$, every fault is detected and this approximates the BJD filter problem. When $s = 1$, only one fault is detected and this approximates the unknown input observer problem.

By using (2) and (3), the dynamic equation of the error, $e = x - \hat{x}$, is

$$\dot{e} = (A - LC)e + \sum_{i=1}^{q} F_i \mu_i + B_w w - Lv$$

Then, the error can be written as

$$e(t) = \Phi(t, t_0)e(t_0) + \int_{t_0}^{t} \Phi(t, \tau)\left(\sum_{i=1}^{q} F_i \mu_i + B_w w - Lv\right) \mathrm{d}\tau \tag{5}$$

subject to

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t, t_0) = (A - LC)\Phi(t, t_0), \quad \Phi(t_0, t_0) = I \tag{6}$$

The residual (4) can be written as

$$r = Ce + v$$

To formulate the robust multiple-fault detection filter problem, it is assumed that $\mu_1 \cdots \mu_q$, w and v are zero mean, white Gaussian noise with power spectral density of $Q_1 \cdots Q_q$, Q_w and V, respectively, and the initial state $x(t_0)$ is a random vector with variance of P_0. It is also assumed that $\mu_1 \cdots \mu_q$, w and v are uncorrelated with each other and with $x(t_0)$. Now a cost criterion is needed for deriving L and $\hat{H}_1 \cdots \hat{H}_s$. If the cost criterion is associated with the projected residual $\hat{H}_i(Ce + v)$, it is unusable from the statistical viewpoint since the variance of the projected residual generates a δ-function due to the sensor noise. Therefore, the cost criterion will be associated with the projected output error $\hat{H}_i Ce$. In order to determine the cost criterion, define

$$h_i(t) \triangleq \hat{H}_i C \int_{t_0}^{t} \Phi(t, \tau) F_i \mu_i \, \mathrm{d}\tau \tag{7a}$$

$$\hat{h}_i(t) \triangleq \hat{H}_i C \int_{t_0}^{t} \Phi(t, \tau) \hat{F}_i \hat{\mu}_i \, \mathrm{d}\tau \tag{7b}$$

$$\bar{h}_i(t) \triangleq \hat{H}_i C\left[\Phi(t, t_0)e(t_0) + \int_{t_0}^{t} \Phi(t, \tau)(B_w w - Lv) \, \mathrm{d}\tau\right] \tag{7c}$$

where $\hat{F}_i = [F_1 \cdots F_{i-1}\ F_{i+1} \cdots F_q]$. From (5), $E[h_i(t)h_i(t)^T]$ represents the transmission from μ_i to $\hat{H}_i Ce$, $E[\hat{h}_i(t)\hat{h}_i(t)^T]$ represents the transmission from $\hat{\mu}_i$ to $\hat{H}_i Ce$ and $E[\bar{h}_i(t)\bar{h}_i(t)^T]$ represents the transmission from w, v and $e(t_0)$ to $\hat{H}_i Ce$ where $E[\bullet]$ is the expectation operator. Note that the power spectral density of $\hat{\mu}_i$ is $\hat{Q}_i = \text{diag}(Q_1 \cdots Q_{i-1}\ Q_{i+1} \cdots Q_q)$ and $e(t_0)$ is a zero mean random vector with variance of P_0 if $\hat{x}(t_0) = E[x(t_0)]$.

Therefore, the robust multiple-fault detection filter problem is to find the filter gain L and projectors $\hat{H}_1 \cdots \hat{H}_s$ which minimize the cost criterion

$$J = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} \text{tr} \left\{ \sum_{i=1}^{s} \left\{ \frac{1}{\gamma_i} E[\hat{h}_i(t)\hat{h}_i(t)^T] + E[\bar{h}_i(t)\bar{h}_i(t)^T] - E[h_i(t)h_i(t)^T] \right\} \right\} \text{d}t \tag{8}$$

where t_1 is the final time and $\gamma_1 \cdots \gamma_s$ are positive scalars. Making $\gamma_1 \cdots \gamma_s$ small places large weightings on reducing the associated nuisance fault transmissions. The summation is used to sum the s projected output error variances for detecting the s faults. The trace operator forms a scalar cost criterion of the matrix output error variance. Note that the power spectral densities $Q_1 \cdots Q_q$ are considered as design parameters. Since no assumption is made on the fault magnitudes, their white noise representation is a convenience. For each projected output error, Q_i and $(1/\gamma_i)\hat{Q}_i$ represent the weightings on the associated target and nuisance fault transmissions, respectively. When Q_i is larger, the transmission from μ_i is larger. This provides a mechanism to enhance the sensitivity of the projected residuals to their associated target faults. When $(1/\gamma_i)\hat{Q}_i$ is larger, the transmission from $\hat{\mu}_i$ is smaller. However, the power spectral densities Q_w and V, and the variance P_0 can have physical values. When Q_w, V and P_0 are larger, the transmission from the process noise, sensor noise and initial condition error is smaller, respectively.

Since the effect of the process and sensor noises on the residual is explicitly minimized, the filter is robust with respect to these disturbances. Certain types of model uncertainties can also be modelled as additive noises [9,13]. Therefore, the filter can be made robust to these model uncertainties. In Section 4, it is shown that the filter recovers the geometric structure of the RDD filter in the limit as $\gamma_i \to 0$, $i = 1 \cdots s$, and the faults are completely isolated. When it is not in the limit, the filter is an approximate RDD filter and only isolates the faults within approximate unobservable subspaces. This new feature allows the filter to be potentially more robust because of the additional design freedom which allows different degrees of fault isolation.

In Section 3, the robust multiple-fault detection filter problem is first solved with $\hat{H}_1 \cdots \hat{H}_s$ defined *a priori* as the projectors used by the RDD filter [3], i.e.

$$\hat{H}_i : \mathscr{Y} \to \mathscr{Y}, \qquad \text{Ker}\, \hat{H}_i = C\hat{\mathscr{T}}_i, \quad \hat{H}_i = I - C\hat{\mathscr{T}}_i[(C\hat{\mathscr{T}}_i)^T C\hat{\mathscr{T}}_i]^{-1}(C\hat{\mathscr{T}}_i)^T \tag{9}$$

where $C\hat{\mathscr{T}}_i = [C\mathscr{T}_1 \cdots C\mathscr{T}_{i-1}\ C\mathscr{T}_{i+1} \cdots C\mathscr{T}_q]$. For time-invariant systems, $C\mathscr{T}_i = [CA^{\delta_{i,1}} f_{i,1}\ CA^{\delta_{i,2}} f_{i,2} \cdots CA^{\delta_{i,p_i}} f_{i,p_i}]$ where $f_{i,j}$ is the jth column of F_i, $\delta_{i,j}$ is the smallest non-negative integer such that $CA^{\delta_{i,j}} f_{i,j} \neq 0$ and $p_i = \dim F_i$. For time-varying systems, the projector (9) is generalized with $C\mathscr{T}_i = [Cb_{i,1,\delta_{i,1}}\ Cb_{i,2,\delta_{i,2}} \cdots Cb_{i,p_i,\delta_{i,p_i}}]$ [10]. The vectors $b_{i,j,\delta_{i,j}}$, $j = 1 \cdots p_i$, are found from the iteration defined by the Goh transformation [14,15],

$$b_{i,j,k} = A(t)b_{i,j,k-1} - \dot{b}_{i,j,k-1} \quad \text{with } b_{i,j,0} = f_{i,j} \tag{10}$$

$\delta_{i,j}$ is the smallest non-negative integer such that $Cb_{i,j,\delta_{i,j}} \neq 0$ for $t \in [t_0, t_1]$. More details about $\mathscr{T}_i$ and $\hat{\mathscr{T}}_i$ can be found in Section 4.1. In Section 4.2, it is shown that (9) minimizes the cost criterion in the limit. Therefore, (9) is the optimal projector in the limit. In Section 5, the robust

multiple-fault detection filter problem is solved with $\hat{H}_1 \cdots \hat{H}_s$ derived from solving the minimization problem.

3. SOLUTION

In this section, the minimization problem given by (8) is solved. By using (7), the cost criterion rewritten as

$$J = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} \operatorname{tr} \left\{ \sum_{i=1}^{s} \left[\hat{H}_i C \int_{t_0}^{t} \Phi(t,\tau) \left(LVL^T + \frac{1}{\gamma_i} \hat{F}_i \hat{Q}_i \hat{F}_i^T - F_i Q_i F_i^T + B_w Q_w B_w^T \right) \right. \right.$$

$$\left. \left. \times \Phi(t,\tau)^T \, \mathrm{d}\tau \, C^T \hat{H}_i + \hat{H}_i C \Phi(t,t_0) P_0 \Phi(t,t_0)^T C^T \hat{H}_i \right] \right\} \mathrm{d}t$$

is to be minimized with respect to L subject to (6). By adding the zero term $1/(t_1 - t_0) \int_{t_0}^{t_1} \operatorname{tr}\{\sum_{i=1}^{s} \hat{H}_i C\{\Phi(t,t)P_i(t)\Phi(t,t)^T - \Phi(t,t_0)P_i(t_0)\Phi(t,t_0)^T - \int_{t_0}^{t} \mathrm{d}/\mathrm{d}\tau[\Phi(t,\tau)P_i(\tau)\Phi(t,\tau)^T]\mathrm{d}\tau\} C^T \hat{H}_i\} \, \mathrm{d}t$ to J and using (6), the minimization problem can be rewritten as

$$\min_L \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} \operatorname{tr} \left\{ \sum_{i=1}^{s} \left[\hat{H}_i C \int_{t_0}^{t} \Phi(t,\tau)(L - P_i C^T V^{-1}) \, V(L - P_i C^T V^{-1})^T \Phi(t,\tau)^T \, \mathrm{d}\tau \, C^T \hat{H}_i \right] \right\} \mathrm{d}t$$

$$= \min_L \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} \operatorname{tr} \left(\sum_{i=1}^{s} \hat{H}_i C W_i C^T \hat{H}_i \right) \mathrm{d}t \tag{11}$$

subject to

$$\dot{W}_i = (A - LC)W_i + W_i(A - LC)^T + (L - P_i C^T V^{-1}) V (L - P_i C^T V^{-1})^T, \; W_i(t_0) = 0 \tag{12}$$

for $i = 1 \cdots s$ where $W_i \geqslant 0$ and

$$\dot{P}_i = AP_i + P_i A^T - P_i C^T V^{-1} C P_i + \frac{1}{\gamma_i} \hat{F}_i \hat{Q}_i \hat{F}_i^T - F_i Q_i F_i^T + B_w Q_w B_w^T, \; P_i(t_0) = P_0 \tag{13}$$

The term $1/(t_1 - t_0) \int_{t_0}^{t_1} \operatorname{tr}(\sum_{i=1}^{s} \hat{H}_i C P_i C^T \hat{H}_i) \, \mathrm{d}t$ is dropped in (11) because it is fixed with respect to L. However, it will be brought back in Section 5 when the cost criterion is also minimized with respect to $\hat{H}_1 \cdots \hat{H}_s$. Note that (13) is solved independently of L and $\hat{H}_1 \cdots \hat{H}_s$.

The variational Hamiltonian of the minimization problem is defined as

$$\mathscr{H} = \sum_{i=1}^{s} \{\operatorname{tr}(\hat{H}_i C W_i C^T \hat{H}_i) + \operatorname{tr}\{K_i[(A - LC)W_i + W_i(A - LC)^T$$

$$+ (L - P_i C^T V^{-1}) V (L - P_i C^T V^{-1})^T]\}\}$$

where K_i is a continuously differentiable matrix Lagrange multiplier. The first-order necessary conditions [16] imply that the optimal solution for L and the dynamics of K_i are

$$\frac{\partial \mathcal{H}}{\partial L} = \sum_{i=1}^{s} [-2CW_iK_i + 2V(L^* - P_iC^TV^{-1})^TK_i] = 0$$

$$\Rightarrow L^* = \left(\sum_{i=1}^{s} K_i\right)^{-1} \left[\sum_{i=1}^{s} K_i(P_i + W_i)\right] C^TV^{-1} \tag{14}$$

and

$$-\dot{K}_i = \frac{\partial \mathcal{H}}{\partial W_i} = K_i(A - LC) + (A - LC)^TK_i + C^T\hat{H}_iC, \quad K_i(t_1) = 0 \tag{15}$$

where $i = 1 \cdots s$. Therefore, the determination of the filter gain requires the solution to a two-point boundary value problem which includes a set of Lyapunov equations (12) and (15), coupled by (14). An alternative approach is to solve (11) numerically by using the gradient method. However, the global minimum cannot be guaranteed because (11) may not be convex. Note that the filter gain computation can be done off-line so that the filter implementation is as straightforward as the RDD filter. Numerical examples are given in Section 6.

For the infinite-time case, the minimization problem (11) becomes

$$\lim_{t_1 \to \infty} \min_L J = \min_L \operatorname{tr}\left(\sum_{i=1}^{s} \hat{H}_iCW_iC^T\hat{H}_i\right) \tag{16}$$

subject to

$$0 = (A - LC)W_i + W_i(A - LC)^T + (L - P_iC^TV^{-1})V(L - P_iC^TV^{-1})^T \tag{17}$$

for $i = 1 \cdots s$ where $W_i \geqslant 0$ and

$$0 = AP_i + P_iA^T - P_iC^TV^{-1}CP_i + \frac{1}{\gamma_i}\hat{F}_i\hat{Q}_i\hat{F}_i^T - F_iQ_iF_i^T + B_wQ_wB_w^T$$

The optimal solution for L can be derived similarly.

$$L^* = \left(\sum_{i=1}^{s} K_i\right)^{-1} \left[\sum_{i=1}^{s} K_i(P_i + W_i)\right] C^TV^{-1} \tag{18}$$

satisfying (17) and

$$0 = K_i(A - LC) + (A - LC)^TK_i + C^T\hat{H}_iC \tag{19}$$

For the special case where $s = 1$ and μ_i is detected, the minimization problem (11) becomes

$$\min_L \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} \operatorname{tr}\left[\hat{H}_iC \int_{t_0}^{t} \Phi(t,\tau)(L - P_iC^TV^{-1})V(L - P_iC^TV^{-1})^T\Phi(t,\tau)^T \,\mathrm{d}\tau C^T\hat{H}_i\right] \mathrm{d}t$$

The optimal solution for L is

$$L^* = P_iC^TV^{-1} \tag{20}$$

Appendix I

"A Decentralized Fault Detection Filter,"

Walter H. Chung, Jason L. Speyer and Robert H. Chen,

***ASME J. of Dynamic Systems, Measurement, Control*, Vol. 123, 2001.**

Walter H. Chung
Jason L. Speyer
Robert H. Chen

Mechanical and Aerospace Engineering Department, University of California, Los Angeles, 48-121 Engr. IV, Los Angeles, CA 90095
e-mail: speyer@seas.ucla.edu or walter.h.chung@aero.org or chrobert@talus.seas.ucla.edu

A Decentralized Fault Detection Filter[1]

In this paper, we introduce the decentralized fault detection filter, a structure that results from merging decentralized estimation theory with the game theoretic fault detection filter. A decentralized approach may be the ideal way to health monitor large-scale systems, since it decomposes the problem down into (potentially smaller) "local" problems. These local results are then blended into a "global" result that describes the health of the entire system. The benefits of such an approach include added fault tolerance and easy scalability. An example given at the end of the paper demonstrates the use of this filter for a platoon of cars proposed for advanced vehicle control systems.
[DOI: 10.1115/1.1367859]

1 Introduction

Observers are, in many ways, an ideal tool for fault detection and identification (FDI). Failures act as unexpected inputs into a system and, thus, drive the error residual of any observer to nonzero values. With careful selection of the observer gain, these fault-driven residuals can be made to have persistent and distinctive characteristics. In many cases, freedom exists to address other design issues, such as noise sensitivity and parameter robustness. For these reasons, the application of observers to the problem of fault detection and identification has long been an active area of research.

There are two types of observers used for fault detection and identification. The first is known as the *Beard-Jones Fault Detection Filter* [1,2]. This filter has a unique subspace structure in which the reachable subspaces of the modeled faults are restricted to lie within nonoverlapping invariant subspaces that can be made unobservable to a projection on the filter residual. Because of this, simultaneous detection and identification can be achieved. The failure is detected when the projection is nonzero. The failure is identified by the subspace corresponding the nonzero projection.

The second type of FDI observer is known as the *unknown input observer*. In this observer, the set of modeled faults is divided into two groups: the faults to be detected and the faults that are to be ignored. The former is made distinguishable from the latter by constructing an output through which the latter set is unobservable. Detection is then achieved when this output is nonzero and identification is trivial because we are only trying to detect one set of faults in the possible presence of the other. The unknown input observer is clearly less capable than the Beard-Jones filter, but its relatively simple structure allows for easy approximation by optimization methods [3,4].

As both of these approaches have become more refined, applications have begun to be seen in the literature [5,6]. With the advent of applications, however, new issues related to implementation have come to the forefront. In this paper, we will look at some of the challenges inherent to detecting faults in large-scale systems. For such systems, a *decentralized fault detection filter* may be the logical approach to the problem.

The decentralized fault detection filter is the result of combining the game theoretic fault detection filter developed by Chung and Speyer [4] with the decentralized filtering algorithm introduced by Speyer [7] and extended by Willsky et al. [8]. It approximates the actions of an unknown input observer and is formed by combining the estimates of several "local" estimators (each driven by independent measurement sets). For large-scale systems, it simplifies the health montoring problem by decomposing it down into a collection of smaller problems. For some systems like a platoon of cars or a formation of airplanes, its decentralized structure reflects the actual physical structure of the system. A decentralized fault detection filter also introduces scalability for circumstances such as when a car joins the platoon or when an airplane drops out of formation for repairs. It also has built in fault tolerance in that sensors can be checked and validated prior to their measurements being blended into the global estimate [9].

The remainder of the paper is organized as follows. In Section 2, the decentralized estimator is described. An essential insight revealed there is that observers that take their gains from Riccati solutions are much more suited for decentralized estimation than general Luenberger Observers that do not. This leads us to a decentralized fault detection filter based upon approximate unknown input observers. We describe these observers in Section 3. An overview of the decentralized fault detection filter is then given in Section 4. An essential part of this filter is how one obtains the global/local decomposition needed to develop the network. We suggest a technique based upon minimal realizations and demonstrate this in Section 5 in an example problem based around a two car platoon.

2 Decentralized Estimation Theory and its Application to FDI

2.1 The General Solution. In this section, we will review the basic results of decentralized estimation theory. A detailed examination of this theory is given in [10]. We begin with a linear system driven by process disturbances, w, and sensor noise, v:

$$\dot{x}=Ax+Bw, \quad x(0),x\in\mathcal{R}^n, \tag{1}$$

$$y=Cx+v, \quad y\in\mathcal{R}^m. \tag{2}$$

It is desired to derive an estimate of x. The standard approach is a full-order observer,

$$\dot{\hat{x}}=A\hat{x}+L(y-C\hat{x}), \quad \hat{x}(0)=0, \tag{3}$$

which we will refer to as a *centralized estimator*. An alternative to this method is to derive the estimate with a *decentralized estimator* in which $\hat{x}$ is found by combining estimates based upon "local" models,

$$\dot{x}^j=A^jx^j+B^jw^j, \quad x^j\in\mathcal{R}^{n^j}, \quad (j=1\ldots N), \tag{4}$$

$$y^j=E^jx^j+v^j, \quad y^j\in\mathcal{R}^{m^j}, \quad (j=1\ldots N). \tag{5}$$

Together these local models provide an alternate representation of the original system, which is referred to as the "global" system

[1]This work was sponsored by Air Force Office of Scientific Research Grant F49620-00-1-0154, NASA Goddard Space Flight Center NAG5-8694 and California Department of Transportation Agreement No. 65A0013, MOU315.

Contributed by the Dynamic Systems and Control Division for publication in the JOURNAL OF DYNAMIC SYSTEMS, MEASUREMENT, AND CONTROL. Manuscript received by the Dynamic Systems and Control Division March 16, 1999; Associate Editor: S. Fassois.

for purposes of clarification. The vector, x, is likewise called the global state. The number of local systems, N, is bounded above by the number of measurements in the system, i.e., $N \leq m$.

The global/local decomposition is really of only secondary importance. As Chung [10] argues, there are no real restrictions on how one forms the global and local models. The real key to the decentralized estimation algorithm is the relationship between the global set of measurements, y, and the N local sets, y^j. The two basic assumptions are that the local sets are simply segments of the global set,

$$y=\begin{Bmatrix} y^1 \\ y^2 \\ \vdots \\ y^N \end{Bmatrix}, \tag{6}$$

and that the local sets can be described in terms of both the local state *and* the global state. In other words, y^j can be given by (5) or by

$$y^j = C^j x + v^j, \quad (j=1 \ldots N). \tag{7}$$

Equations (2), (6), and (7) imply that

$$C=\begin{bmatrix} C^1 \\ \vdots \\ C^N \end{bmatrix}$$

and that

$$v=\begin{bmatrix} v^1 \\ \vdots \\ v^N \end{bmatrix}. \tag{8}$$

The decentralized estimation algorithm falls out when we attempt to estimate the global state by first generating estimates of the local systems (4) using the local measurement sets, y^j, and the local models, A^j:

$$\dot{\hat{x}}^j = A^j \hat{x}^j + L^j (y^j - E^j \hat{x}^j), \quad \hat{x}^j(t_0)=0, \quad (j=1 \ldots N). \tag{9}$$

The objective then is to obtain the global state estimate, $\hat{x}$, through some simple function of the local estimates. As it turns out, in the most general case, the global estimate is an affine combination,

$$\hat{x} = \sum_{j=1}^{N} (G^j \hat{x}^j + h^j), \tag{10}$$

where h^j is a measurement-dependent variable propagated by

$$\dot{h}^j = \Phi h^j + (\Phi G^j - \dot{G}^j - G^j \Phi^j)\hat{x}^j, \quad h^j(0)=0. \tag{11}$$

The constituent matrices are defined as

$$\Phi := A - \sum_{j=1}^{N} G^j L^j C^j,$$

$$\Phi^j := A^j - L^j E^j.$$

The G^j matrices are "blending matrices." They are so-called because they act to blend the local estimates together to form the global estimate. They can also be shown [10] to directly connect the local and global gains via

$$L=[G^1 \ldots G^N]\begin{bmatrix} L^1 & 0 & \ldots & 0 \\ 0 & L^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & L^N \end{bmatrix}. \tag{12}$$

The interested reader can derive Eqs. (11) and (12) by differentiating (10) and substituting in the equations for the local estimators where appropriate. The derivation is completed through some algebraic manipulation and integration by parts (see [10] or [11]).

Equation (12) looks harmless, but it turns out to be the key relationship in decentralized estimation. In fact, it is the necessary and sufficient condition for decentralized estimation [10,11]. Another interesting fact is that (12) does not have a solution in the general case for the blending matrices, G^j, because of an insufficient number of equations for all of the unknowns. There is, however, one general class of estimator for which (12) is satisfied almost automatically. This class is comprised of estimators that take their gains from Riccati solutions, i.e., Kalman filters [7,8] or H^∞ filters [12]. In this case, the local gains are found from

$$L^j = P^j (E^j)^T (V^j)^{-1}, \tag{13}$$

where, in the case of the Kalman filter, the matrix P^j is the solution of the Riccati equation:

$$\dot{P}^j = A^j P^j + P^j (A^j)^T + B^j W^j (B^j)^T - P^j (E^j)^T (V^j)^{-1} E^j P^j,$$

$$P^j(0) = P_0^j.$$

The matrices, V^j and W^j, are weightings that are related to the local disturbances, v^j and w^j, that drive the local systems (4), (5). For the Kalman filter, it is assumed that v^j and w^j are white,

$$E[w^j(t) w^j(\tau)^T] = W^j \delta(t-\tau)$$

$$E[v^j(t) v^j(\tau)^T] = V^j \delta(t-\tau),$$

and Gaussian. The initial condition, P_0^j, is chosen by the analyst based upon his knowledge of the system. In the global system, the gain is

$$L = P C^T \mathcal{V}^{-1},$$

where

$$\mathcal{V}=\begin{bmatrix} V^1 & 0 & \cdots & 0 \\ 0 & V^2 & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & \cdots & V^N \end{bmatrix}. \tag{14}$$

The matrices, V^j, on the block diagonal of $\mathcal{V}$ are the local measurement noise weightings. In our example, however, we will show that there is some design flexibility in choosing the global weight. Specifically, one can choose scalar gains on the local weightings,

$$\mathcal{V}=\begin{bmatrix} \alpha_1 V^1 & 0 & \cdots & 0 \\ 0 & \alpha_2 V^2 & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & \cdots & \alpha_N V^N \end{bmatrix}. \tag{15}$$

This added flexibility allows us to meet other design criteria that might arise in the problem. In our example in Section 5, we demonstrate how to use this design freedom to improve the response of our decentralized fault detection filter to the faults that we want to see.

The matrix, P, is the solution to the global Riccati equation,

$$\dot{P} = AP + PA^T + B\mathcal{W}B^T - PC^T\mathcal{V}CP, \quad P(0)=P_0.$$

The blending matrix solution is then

$$G^j = P(S^j)^T (\alpha_j P^j)^{-1} \quad j=1,\ldots,N, \tag{16}$$

where S^j is any matrix such that

$$C^j = E^j S^j. \tag{17}$$

One can, in fact, always take $S^j = (E^j)^\dagger C^j$ where $(E^j)^\dagger$ is the pseudo-inverse of E^j [8]. Note that the solutions for G^j will always exist for Riccati-based observers so long as P^j is invertible

or, equivalently, positive-definite. This will always be the case if the triples, (C^j,A^j,B^j), are controllable and observable for each of the local systems.

2.2 Implications for Detection Filters. The analysis of the previous section implies that we will be able to form a decentralized fault detection filter in the general case only if we are able to find a Riccati-based observer that is equivalent to a Beard-Jones filter or unknown input observer. The most direct way to achieve this is to find a linear-quadratic optimization problem that is equivalent to the fault detection and identification problem. This is an analog of the famous inverse optimal control problem first posed by Kalman [13]. In [4], however, it is shown that the Beard-Jones filter gains do not correspond with those derived from linear-quadratic problems. An indirect way to get a Riccati-based observer is to pose a linear-quadratic optimization problem that closely mimics the fault detection problem. Such a problem was posed and solved in [4], and we will review the solution found there in the next section.

3 The Approximate Fault Detection and Identification Problem

3.1 Problem Formulation. Consider the system given by (1), (2) with the further assumption that the state matrices have sufficient smoothness to guarantee the existence of derivatives various order. Beard [14] showed that failures in the sensors and actuators, and unexpected changes in the plant dynamics can be modeled as additive signals,

$$\dot{x}=Ax+Bw+F_1\mu_1+\cdots+F_q\mu_q. \tag{18}$$

Let n be the dimension of the state-space. The $n\times p_i$ matrix, F_i, $i=1\cdots q$, is called a failure map and represents the directional characteristics of the ith fault. The $p_i\times 1$ vector, μ_i, is the failure signal and represents the time dependence of the failure. It will always be assumed that each F_i is monic, i.e., $F_i\mu_i\neq 0$ for $\mu_i\neq 0$. See [15,4] for further details on how to model failures. Throughout this paper, we will refer to μ_1 as the "target fault" and the other faults, μ_j, $j=2\cdots q$, as the "nuisance faults." Without loss of generality, we can represent the entire set of nuisance faults (and, if desired, the disturbance, w) with a single map, F_2, and vector, μ_2:

$$\dot{x}=Ax+F_1\mu_1+F_2\mu_2.$$

Suppose that it is desired to detect the occurrence of the failure, μ_1, in spite of the measurement noise, v, and the possible presence of the nuisance faults, μ_2. The Beard-Jones filter solves this problem by finding a gain, L, so that a standard Luenberger Observer,

$$\dot{\hat{x}}=A\hat{x}+L(y-C\hat{x}), \tag{19}$$

will have an invariant subspace structure that restricts the influence of μ_1 and μ_2 to separate and nonintersecting invariant subspaces. With a properly chosen projector, H, we can then project the filter residual, $(y-C\hat{x})$, onto the orthogonal complement of the invariant subspace containing μ_2 and get a signal,

$$z=H(y-C\hat{x}), \tag{20}$$

such that

$$z=0 \quad \text{when } \mu_1=0 \ (\mu_2 \text{ is arbitrary}). \tag{21}$$

To be useful for FDI, z must also be such that

$$z\neq 0 \quad \text{when } \mu_1\neq 0. \tag{22}$$

If we restrict ourselves to time-invariant systems, (22) will be equivalent to requiring that the transfer function matrix between $\mu_1(t)$ and $z(t)$ to be left-invertible. Left-invertibility, however, is a severe restriction, and it has no analog for the general time-varying systems that we want to consider here. Previous researchers [15,16] have, in fact, only required that the mapping from $\mu_1(t)$ to $z(t)$ be input observable, i.e., $z\neq 0$ for any μ_1 that is a step input. It can be argued [16] that z will be nonzero for "almost any" μ_1, since μ_1 is unlikely to remain in the kernal of the mapping to z for all time.

We formulate the approximate detection filter design problem by requiring input observability and relaxing the requirement (21). Instead of (21), we require only that the transmission of the nuisance fault be bounded above by a preset level, $\gamma>0$:

$$\frac{\|z\|^2}{\|\mu_2\|^2}\leq\gamma. \tag{23}$$

Equation (23) is identical to the disturbance attenuation problem from robust control theory. We refer to the solution to the approximate detection filter problem as the *game theoretic fault detection filter.*

We complete our formulation of the disturbance attenuation problem for fault detection by constructing the projector, H, that determines the failure signal, z. For time-invariant systems, this projector is constructed to map the invariant subspace containing the range of F_2 to zero [14,15], i.e.,

$$H=I-C\hat{F}[(C\hat{F})^TC\hat{F}]^{-1}(C\hat{F})^T, \tag{24}$$

where

$$\hat{F}=[A^{\beta_1}f_1,\ldots,A^{\beta_{p_2}}f_{p_2}]. \tag{25}$$

The vector f_i, $i=1\cdots p_2$, is the ith column of F_2, and the integer β_i is the smallest natural number such that $CA^{\beta_i}f_i\neq 0$. With little additional effort, this result can be extended to the time-varying case,

$$H=I-C\hat{F}(t)[(C\hat{F}(t))^TC\hat{F}(t)]^{-1}(C\hat{F}(t))^T. \tag{26}$$

The columns of the matrix,

$$\hat{F}(t)=[b_1^{\beta_1}(t),\ldots,b_{p_2}^{\beta_{p_2}}(t)], \tag{27}$$

are constructed with the Goh Transformation [4]:

$$b_i^1(t)=f_i(t), \tag{28}$$

$$b_i^j(t)=A(t)b_i^{j-1}(t)-\dot{b}_i^{j-1}. \tag{29}$$

In the time-varying case, β_i is the smallest integer for which the interaction above leads to a vector $b_i^{\beta_k}(t)$ such that $C(t)b_i^{\beta_k}(t)\neq 0$ for all $t\in[t_0,t_1]$. It will be assumed that $A(t)$, $C(t)$, and $F_2(t)$ are such that β_i exists. Since the state-space has dimension, n, β_i is such that $0\leq\beta_i\leq n-1$.

Remark 1. One of the advantages to the disturbance attenuation approach to designing FDI Observers is that the time-varying case can be handled as easily as the time-invariant case. This is an improvement over classic detection filter designs.

We are now ready to discuss the conditions under which the solution to (23) will also generate an input observable mapping from μ_1 to z. The key requirement is that the system be *output separable.* That is, F_1 and F_2 must be linearly independent and remain so when mapped to the output space by C and A. For time-invariant systems, the test for output separability is

$$\text{rank } [CA^{\delta_1}\bar{f}_1,\ldots,CA^{\delta_{p_1}}\bar{f}_{p_1},CA^{\beta_1}f_1,\ldots,CA^{\beta_{p_2}}f_{p_2}] = p_1+p_2. \tag{30}$$

As in (25), f_i is the ith column of F_2, and β_i is the the smallest integer such that $CA^{\beta_i}f_i\neq 0$. Similarly, $\bar{f}_j$ is the jth column of F_1, and δ_j is the smallest integer such that $CA^{\delta_j}\bar{f}_j\neq 0$. The integer sum, p_1+p_2, is the total number of columns in F_1 and F_2.

For time-varying systems, the output separability test becomes

$$\text{rank } [C(t)\bar{b}_1^{\delta_1}(t),\ldots,C(t)\bar{b}_{p_1}^{\delta_{p_1}}(t),C(t)b_1^{\beta_1}(t),\ldots,$$

$$C(t)b_{p_2}^{\beta_{p_2}}(t)]=p_1+p_2, \quad \forall t\in[t_0,t_1], \tag{31}$$

where the vectors, $b_i^{\beta_i}$ and $\bar{b}_j^{\delta_j}$, are found from the iteration defined by (28) and (29). The initial vector, $\bar{b}_j^1$, is set equal to the jth column of F_1, and b_i^1 is initialized as the ith column of F_2.

The following proposition, given in [4], connects output separability to input observability and shows the importance of the monicity assumption:

Theorem 2. Suppose that a given filter satisfies (23) and generates the failure signal z given by (20). If F_1 and F_2 are output separable and F_1 is monic, then the mapping, $\mu_1(t)\rightarrow z(t)$, is input observable.

3.2 A Game Theoretic Solution. We now turn our attention to the disturbance attenuation problem implied by (23). We begin by defining a disturbance attenuation function,

$$D_{af}=\frac{\int_{t_0}^{t_1}\|HC(x-\hat{x})\|_Q^2 dt}{\int_{t_0}^{t_1}[\|\mu_2\|_{M^{-1}}^2+\|v\|_{V^{-1}}^2]dt+\|x(t_0)-\hat{x}_0\|_{P_0}^2}. \tag{32}$$

D_{af} is simply a ratio of the outputs over the disturbances. Equation (32) is patterned roughly after (23). We have added the sensor noise, v, and the initial error, $x(t_0)-\hat{x}_0$, to the set of disturbance signals to inject tradeoffs for noise rejection and settling time into the problem. M, V, Q, and P_0 are weighting matrices. Note that we do not include the target fault, μ_1, at this stage of the design problem, since we are now focusing on nuisance blocking. Our only concern with μ_1 is that it be visible at the output, which is what Theorem 2 guarantees. The disturbance attenuation problem is to find the estimate, $\hat{x}$, so that for all μ_2, $v\in L_2[t_1,t_2]$, and $x(t_0)\in\mathcal{R}^n$,

$$D_{af}\leqslant\gamma.$$

The positive real number, γ, is called the disturbance attenuation bound. (C,A) will always be assumed to be an observable pair.

To solve this problem, we convert (32) into a cost function,

$$J=\int_{t_0}^{t_1}[\|HC(x-\hat{x})\|_Q^2-\gamma(\|\mu_2\|_{M^{-1}}^2+\|y-Cx\|_{V^{-1}}^2)]dt$$
$$-\|x(t_0)-\hat{x}_0\|_{\Pi_0}^2, \tag{33}$$

where we have used (2) to rewrite the measurement noise term. Note that we have also rewritten the initial error weighting, defining $\Pi_0=\gamma P_0$. The disturbance attenuation problem is then solved via the differential game,

$$\min_{\hat{x}}\max_{y}\max_{\mu_2}\max_{x(t_0)}J\leqslant 0, \tag{34}$$

subject to

$$\dot{x}=Ax+F_2\mu_2,$$
$$y=Cx+v. \tag{35}$$

Those familiar with linear-quadratic optimization, will recognize the solution of the differential game [4] to be a Luenberger Observer,

$$\dot{\hat{x}}=A\hat{x}+\gamma\Pi^{-1}C^TV^{-1}(y-C\hat{x}), \quad \hat{x}(t_0)=\hat{x}_0, \tag{36}$$

whose gain is taken from the solution to a Riccati equation,

$$-\dot{\Pi}=A^T\Pi+\Pi A+\frac{1}{\gamma}\Pi F_2MF_2^T\Pi$$
$$+C^T(HQH-\gamma V^{-1})C \quad \Pi(t_0)=\Pi_0. \tag{37}$$

In many cases, it is desired to extend finite-time solutions of game theoretic problems to the steady-state condition. Whenever it is possible to find such a solution, the optimal estimator will be given by (36) with Π being the solution of the algebraic Riccati equation,

$$0=A^T\Pi+\Pi A+\frac{1}{\gamma}\Pi F_2MF_2^T\Pi+C^T(HQH-\gamma V^{-1})C. \tag{38}$$

However, unlike linear quadratic optimal control problems, there are no conditions which guarantee the existence of a unique, nonnegative definite, stabilizing solution to the steady-state Riccati equation, except in the special case where A is asymptotically stable [17].

4 The Decentralized Fault Detection Filter

Given the results of the previous two sections, we now propose a decentralized fault detection filtering algorithm. The essential idea is to implement the Riccati-based game theoretic fault detection filter as a decentralized estimator. An overview of the procedure is as follows:

1 Identify the sensors and actuators which must be monitored at the global level, i.e., define the target faults for the global filter.

2 Identify the faults that should be included in the global nuisance set. The remaining faults should be monitored at the local levels.

3 Derive global and local models for the system including failure maps. Chung [4] contains a brief discussion about this process. We will demonstrate one method in which the local models are derived from the global model via a minimum realization.

4 Design game theoretic fault detection filters for the local and global systems. Solve the corresponding Riccati equations and store the solutions for later use.

5 Determine the blending solutions G^j from Eq. (16).

6 Propagate the local estimates, $\hat{x}^j$, and vectors, h^j, and then use the decentralized estimation algorithm (10) to derive a global estimate, $\hat{x}$.

7 Determine the global failure signal from $(y-C\hat{x})$ where y is the total measurement set, C is the global measurement matrix, and $\hat{x}$ is the global fault detection filter estimate just derived.

Remark 3. Minimum realizations leave only those states that are both observable and controllable. Our use of minimum realizations in step #3 extracts the local models from the global model by pulling out only those states (or combinations of states) that are observable through the local measurements, y^j, and driven by the failures chosen to be included in the local model. Determining a compatible and consistent local/global decomposition is a key issue in decentralized estimation and control. The use of minimum realizations that we suggest here φ is a logical and theoretical rigorous approach to this problem.

5 Range Sensor Fault Detection in a Platoon of Cars

5.1 Problem Statement. We will now examine the utility of the decentralized approach to FDI by working through an example. The problem that we will look at involves the detection of failures within a system of two cars traveling as a platoon (see Fig. 1). The cars are controlled to maintain a uniform speed and constant separation. The platoon is the central component of automated highway schemes in which groups of cars line up single file and travel as a unit. The objective is to eliminate the backup caused by the interaction of individual vehicles maneuvering across highway lanes [18,19]. The viability of the platooning scheme, however, will depend on many factors, not the least of which are reliability and safety.

The FDI schemes that we have examined to this point are capable of monitoring individual cars, but may not be ideal for monitoring elements that deal with the interactions between cars. For example, to maintain uniform speed throughout the platoon and to keep the spacing between the cars constant, additional sen-

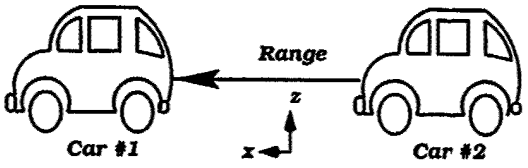


Fig. 1 Two-car platoon with range sensor

sors will be needed to measure the relative speed and the relative distance, or "range," between the cars. In order to detect a failure in the range sensor using analytical redundancy, however, it is necessary to have a dynamical relationship between the range sensor and other sensors on the vehicles. Range, however, involves the dynamics of both of the cars and so would require a higher-order model for its detection filter.

While this is not necessarily prohibitive, it does not make use of the many different state estimates that are already being propagated throughout the platoon. The sensors on each of the cars, for instance, will be monitored by detection filters, and it is more than likely that a state estimate would also be generated by the vehicles' control loops. Given these pre-existing estimates, it seems logical to make use of the decentralized estimation algorithm to carry out range sensor fault detection.

The presentation of the example is as follows. In the next subsection, we present the problem and the model of a single car derived in [18]. We then manipulate this model into a two car platoon model and define the target and nuisance faults. Referring back to the steps listed in Section 4, these are steps #1, #2, and part of #3. In Section 5.3, we complete step #3 by deriving the local models from the global one. In Section 5.4, we design game theoretic filters for the local and global problems and calculate the blending matrices (steps #4 and #5). We also implement the decentralized estimator equations (step #6) and monitor the generated residual for indications of a Range sensor failure (step #7).

5.2 System Dynamics and Failure Modeling. Our example starts with the car model used in [18]. In this model, the nonlinear, six degree-of-freedom dynamics of a representative automobile are linearized about a straight, level path at a speed of 25 meters/s (roughly 56 miles per hour). The linearized equations are found to decouple nicely into lateral and longitudinal dynamics, much like an airplane. Moreover, the linearized equations can be further reduced by eliminating "fast modes" and actuator states. For simplicity, we will only use the longitudinal dynamics which we represent as

$$\dot{x}=A^L x,$$

$$y=C^L x,$$

where the superscript "L" stands for "longitudinal." The vehicle states are

$$x=\begin{Bmatrix} m_a \\ \omega_e \\ v_x \\ v_z \\ z \\ q \\ \theta \end{Bmatrix} \begin{matrix} \text{engine air mass (kg)} \\ \text{engine speed (rad/s)} \\ \text{long. velocity(m/s)} \\ \text{vertical velocity (m/s)} \\ \text{vertical position (m)} \\ \text{pitch rate (rad/s)} \\ \text{pitch (rad)} \end{matrix} \tag{39}$$

and are propagated by the state matrix,

$$A^L=\begin{bmatrix} -22.56 & -0.11683 & 0 & 0 & 0 & 0 & 0 \\ 307.03 & -35.412 & 397.43 & -238.06 & -2698 & -3753 & -331.14 \\ 0 & 0.071298 & -0.81773 & 0.59338 & 6.7786 & 16.807 & 1.5162 \\ 0 & -0.0019628 & 0.022119 & -3.5646 & -40.421 & -9.0765 & -0.81415 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & -0.019628 & 0.22118 & -0.61304 & -7.1619 & -39.926 & -3.6293 \end{bmatrix}. \tag{40}$$

The measurements are

$$y=\begin{Bmatrix} m_a \\ \omega_e \\ \dot{v}_x \\ \dot{v}_z \\ q \\ \bar{\omega}_f \\ \bar{\omega}_r \end{Bmatrix} \begin{matrix} \text{engine air mass (kg)} \\ \text{engine speed (rad/s)} \\ \text{long. acceleration (m/s}^2\text{)} \\ \text{vertical acceleration (m/s}^2\text{)} \\ \text{pitch rate (rad/s)} \\ \text{front symmertric wheel speed (rad/s)} \\ \text{rear symmertric wheel speed (rad/s)} \end{matrix} \tag{41}$$

with the corresponding measurement matrix,

$$C^L=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0713 & -0.8177 & 0.5934 & 6.7786 & 16.8068 & 1.5162 \\ 0 & -0.0020 & 0.0221 & -3.5646 & -40.4210 & -9.0765 & -0.8141 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 7.1220 & -4.5806 & -51.9152 & 58.8718 & 5.1944 \\ 0 & 0.0888 & 5.9738 & -3.5782 & -40.5542 & -56.4109 & -4.9773 \end{bmatrix}. \tag{42}$$

The rear and front symmetric wheel speeds are states that were eliminated when the fast modes were factored out of the linearized system.

5.3 Global and Local Decomposition. In order to build a detection filter for the range sensor, we need to use (39)–(42) to build state space models for the platoon,

$$\dot{\eta}=A\eta+F_1\mu_1+F_2\mu_2,$$

$$y=C\eta,$$

and the two individual cars,

$$\dot{\eta}^1=A^1\eta^1+F_1^1\mu_1^1+F_2^1\mu_2^1,$$

$$y^1=E^1\eta^1,$$

$$\dot{\eta}^2=A^2\eta^2+F_1^2\mu_1^2+F_2^2\mu_2^2,$$

$$y^2=E^2\eta^2.$$

We will build up our models with the following steps:

1 Using (39)–(42), we will derive the global state matrices, A and C.
2 Using the modelling techniques described in [15] and [4], we will determine the failure maps, F_i.
3 We will then obtain the local state matrices, A^i, E^i, and F_j^i, from the minimum realization of the triples (C^1,A,F_2) and (C^2,A,F_2).

The obvious way to get the global matrices, A and C, is to form block diagonal composite matrices with A^L and C^L repeated on the diagonal, i.e.,

$$A'=\begin{bmatrix}A^L & 0\\ 0 & A^L\end{bmatrix},\quad C'=\begin{bmatrix}C^L & 0\\ 0 & C^L\end{bmatrix}.$$

This, however, is not sufficient, since there is no way to describe the range, R, between the two vehicles with the given states (39). Range is the relative distance between the cars,

$$R=x^1-x^2,$$

where x^i is the longitudinal displacement of car i. Displacement, however, is not a state of the vehicle (39). We must, therefore, add a range state to the platoon dynamics, using the equation,

$$\dot{R}=v_x^1-v_x^2.$$

The end result is that the platoon will be a fifteen-state system,

$$\eta=\begin{Bmatrix}m_a^1\\ \omega_e^1\\ v_x^1\\ v_z^1\\ z^1\\ q^1\\ \theta^1\\ m_a^2\\ \omega_e^2\\ v_x^2\\ v_z^2\\ z^2\\ q^2\\ \theta^2\\ R\end{Bmatrix}\begin{matrix}\text{engine air mass (kg)—Car\#1}\\ \text{engine speed (rad/s)—Car\#1}\\ \text{long. velocity (m/s)—Car\#1}\\ \text{vertical velocity (m/s)—Car\#1}\\ \text{vertical position (m)—Car\#1}\\ \text{pitch rate (rad/s)—Car\#1}\\ \text{pitch (rad)—Car\#1}\\ \text{engine air mass (kg)—Car\#2}\\ \text{engine speed (rad/s)—Car\#2}\\ \text{long. velocity (m/s)—Car\#2}\\ \text{vertical velocity (m/s)—Car\#2}\\ \text{vertical position(m)-Car\#2}\\ \text{pitch rate (rad/s)—Car\#2}\\ \text{pitch (rad)—Car\#2}\\ \text{Range (m).}\end{matrix}$$

The corresponding state matrix is

$$A=\begin{bmatrix}A^L & 0 & 0\\ 0 & A^L & 0\\ E_1 & -E_1 & 0\end{bmatrix},\tag{43}$$

$$E_1=[0\,0\,1\,0\,0\,0\,0].$$

The measurement matrix is

$$C=\left[\begin{array}{ccc}C^L & & 0\\ \hline & C^L & 0\\ 0 & 0 & 1\end{array}\right]=\begin{bmatrix}C^1\\ C^2\end{bmatrix},\tag{44}$$

where C^1 and C^2 can be inferred from (44). Finally, the local measurement sets are

$$y^1=\begin{Bmatrix}m_a^1\\ \omega_e^1\\ \dot{v}_x^1\\ \dot{v}_z^1\\ q^1\\ \bar{\omega}_f^1\\ \bar{\omega}_r^1\end{Bmatrix}\begin{matrix}\text{engine air mass (kg)—Car\#1}\\ \text{engine speed (rad/s)—Car\#1}\\ \text{long. acceleration (m/s}^2\text{)—Car\#1}\\ \text{vertial acceleration (m/s}^2\text{)—Car\#1}\\ \text{pitch rate (rad/s)—Car\#1}\\ \text{front symmertric wheel speed (rad/s)—Car\#1}\\ \text{rear symmetric wheel speed (rad/s)—Car\#1.}\end{matrix}$$

and

$$y^2=\begin{Bmatrix}m_a^2\\ \omega_e^2\\ \dot{v}_x^2\\ \dot{v}_z^2\\ q^2\\ \bar{\omega}_f^2\\ \bar{\omega}_r^2\\ R\end{Bmatrix}\begin{matrix}\text{engine air mass (kg)—Car\#2}\\ \text{engine speed (rad/s)—Car\#2}\\ \text{long. acceleration (m/s}^2\text{)—Car\#2}\\ \text{vertial acceleration (m/s}^2\text{)—Car\#2}\\ \text{pitch rate (rad/s)—Car\#2}\\ \text{front symmertric wheel speed (rad/s)—Car\#2}\\ \text{rear symmetric wheels peed (rad/s)—Car\#2}\\ \text{range (m).}\end{matrix}$$

Our ultimate objective is to design a filter that will detect a range sensor fault in the presence of potential failures in the other sensors. In an actual health monitoring system, we would design the global filter to block out all of the nuisance faults that are output separable from the range sensor fault and then rely upon the local filters to monitor the remaining faults. Given the size of our example, however, the full analysis required to do a detailed design would clutter our presentation. We will, therefore, limit ourselves to constructing only one local filter per car and will choose simple nuisance sets at both the global and local levels.

For this example, we choose to monitor the front symmetric wheel speed sensor at the local level. The nuisance set is then chosen to be the engine air mass sensor and the vertical accelerometer. At the global level, the range sensor has already been designated as the target fault. We, therefore, complete the problem definition by choosing the engine speed sensor and longitudinal accelerometer as the global nuisance set. There is no particular significance attached to any of our choices for the nuisance and target sets, aside from the choice of the range sensor as the global target fault.

Following standard modeling techniques [15,4], we construct the two engine speed sensor failure maps $F_{\omega_e^1}$ and $F_{\omega_e^2}$. To save space we do not list these matrices out explicitly. The interested reader can refer to [11]. To complete the problem, we also need to construct maps for the accelerometer failures, $F_{\dot{v}_x^1}$ and $F_{\dot{v}_x^2}$, and the range sensor, F_R. For the local filters, failure maps need to be constructed for the airmass sensors, $F_{m_a^1}$ and $F_{m_a^2}$, vertical accelerometers, $F_{\dot{v}_z^1}$ and $F_{\dot{v}_z^2}$, and front wheel speed sensors, $F_{\bar{\omega}_f^1}$ and $F_{\bar{\omega}_f^2}$. A quick application of (30) will show that all of our failure sets are output separable.

We are now in position to generate the local state equations. The local dynamics for car #1 come from the minimum realization of $(C^1, A, [F_{m_a^1} F_{\dot{v}_z^1}])$. The corresponding matrices are

$$A^1 = \begin{bmatrix} -0.087694 & 0.0038094 & -0.12133 & -0.010701 & 3.9941 & 42.617 & 1.2879 \\ 0.032194 & 1.6765 & 57.123 & 7.2346 & 26.27 & -665.78 & 496.6 \\ 0.00005 & -0.021736 & -22.56 & 0.11478 & -0.0001 & 0.00008 & -0.00005 \\ -0.077512 & 7.7689 & -301.66 & -38.647 & -137.16 & 3612 & -2816.7 \\ -0.096212 & -0.073026 & 2.498 & 0.2312 & 0.89067 & -19.054 & 9.0737 \\ -0.94943 & -0.26102 & -0.20407 & -0.067025 & -0.41229 & -2.4689 & 0.16425 \\ -0.27186 & 0.092418 & 0.12024 & 0.19024 & -0.010912 & -1.302 & -1.434 \end{bmatrix},$$

$$E^1 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -0.0004 & 0.18605 & 0 & -0.98251 & 0.008136 & -0.000665 & 0.000392 \\ 0.0043561 & -0.014182 & 0 & -0.090334 & -0.2118 & 11.266 & -14.31 \\ 0.00015951 & -0.00067636 & 0 & -0.0048006 & -4.0642 & -41.318 & -2.4264 \\ -0.00014266 & -0.97872 & 0 & -0.18537 & 0.0016064 & 0.024547 & -0.084511 \\ -0.00030256 & 0.0016942 & 0 & 0.0069288 & 1.4478 & -34.102 & -71.377 \\ 0.0009564 & -0.0038718 & 0 & -0.019192 & 2.1041 & -55.207 & 42.987 \end{bmatrix},$$

$$F^1_{m_a^1} = \begin{bmatrix} 0 & -0.12133 \\ 0 & 57.1230 \\ 1 & -22.5605 \\ 0 & -301.6586 \\ 0 & 2.4980 \\ 0 & -0.2041 \\ 0 & 0.12024 \end{bmatrix}, \quad F^1_{\dot{v}_z^1} = \begin{bmatrix} 7.9031 & -1.6879 \\ -0.0007 & -0.0213 \\ 0 & 0 \\ -0.0048 & -0.0057 \\ -0.1760 & -0.7911 \\ -0.0068 & -7.4136 \\ -0.0003 & -2.1388 \end{bmatrix}.$$

The model for Car #2 is similarly found by obtaining the minimum realization of $(C^2, A, [F_{m_a^2} F_{\dot{v}_z^2}])$. The corresponding matrices are

$$A^2 = \begin{bmatrix} -0.26387 & -0.27372 & 0.97419 & -0.040683 & 0 & 0 & 0 & 0 \\ 0.28256 & 0.2607 & 0.042752 & 1.0237 & 0 & 0 & 0 & 0 \\ -12.546 & -12.054 & -1.4539 & -0.79488 & -0.002510 & 0.0001643 & 0.000136 & 0.03416 \\ -28.279 & -27.514 & -2.1059 & -3.0468 & 0.004805 & 8.129 & 6.711 & -0.06539 \\ 195.07 & 193.92 & -2.3745 & 38.898 & -0.19848 & -152.87 & -126.21 & 2.7044 \\ 3.8593 & 4.598 & 0.3571 & 0.5193 & -0.000005 & -21.332 & -18.419 & 0.00006 \\ -4.0915 & -4.8456 & -0.37617 & -0.5471 & 0.00003 & 22.827 & 17.824 & -0.00042 \\ -2654.8 & -2639.1 & 32.315 & -529.37 & -304.44 & 2080.5 & 1717.5 & -57.774 \end{bmatrix},$$

$$E^2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0.9973 & 0 & 0 & 0.0733 \\ 0 & 0 & 0 & 0 & 0.0733 & 0 & 0 & -0.9973 \\ -12.008 & -11.76 & -0.5409 & -1.6668 & 0.00522 & 5.2402 & 4.3261 & -0.0711 \\ 5.9034 & 6.9535 & 0.5329 & 0.7915 & -0.00014 & -31.362 & -25.727 & 0.00196 \\ -0.0112 & 0.01132 & -0.7316 & -0.6816 & 0 & -0.00195 & -0.00161 & 0 \\ -43.291 & -39.922 & -8.6999 & 1.9601 & 0 & -40.162 & -33.156 & 0 \\ 40.011 & 39.775 & -0.4870 & 7.9783 & 0.0065 & -31.356 & -25.886 & -0.08855 \\ 0.69369 & -0.71973 & -0.01465 & -0.007574 & 0 & -0.01755 & -0.01449 & 0 \end{bmatrix},$$

$$F^2_{m_a^2}=\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0.9973 & 0.0002 \\ 0 & 0 \\ 0 & 0 \\ 0.0733 & -307.8575 \end{bmatrix}, \quad F^2_{\dot{v}_z^2}=\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ -5.0327 & -4.9282 \\ 6.0961 & -6.2254 \\ 0 & 0 \end{bmatrix}.$$

With all of these system matrices in place, we can now form the residual projectors, H, needed generate the failure signal, z. In the global filter, we define

$$\hat{F}=[F_{\omega_e^1}F_{\dot{v}_x^1}F_{\omega_e^2}F_{\dot{v}_x^2}].$$

In the local filters, we define

$$\hat{F}^i=[F^i_{m_a^i}F^i_{\dot{v}_z^i}] \quad i=1,2$$

The projectors, H and H^i, are then found by applying (24). Again, we do not show either of these matrices to save space.

5.4 Decentralized Fault Detection Filter Design. We will first design filters for the local systems. As with all Riccati-based filters, the central step in the process is in obtaining a solution to the appropriate Riccati equation. For simplicity, we will use the steady-state version. Typically, one iterates on the design by trying various combinations of weightings until a Riccati solution is found that leads to a filter that gives the best tradeoff between target fault transmission and nuisance fault attenuation. For this example, it was found that choosing

$$M^1=10\times I_4, \quad V^1=I_7,$$

$$Q^1=I_7, \quad \gamma_1=0.18$$

as the weightings and attenuation bound leads to a filter gain,

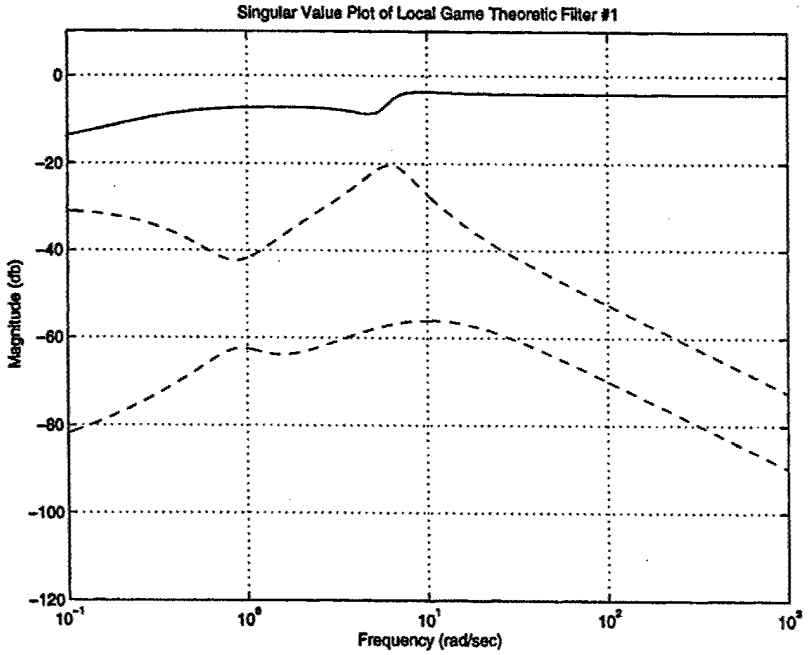


Fig. 2 Platoon example—signal transmission in local detection filter on car # 1

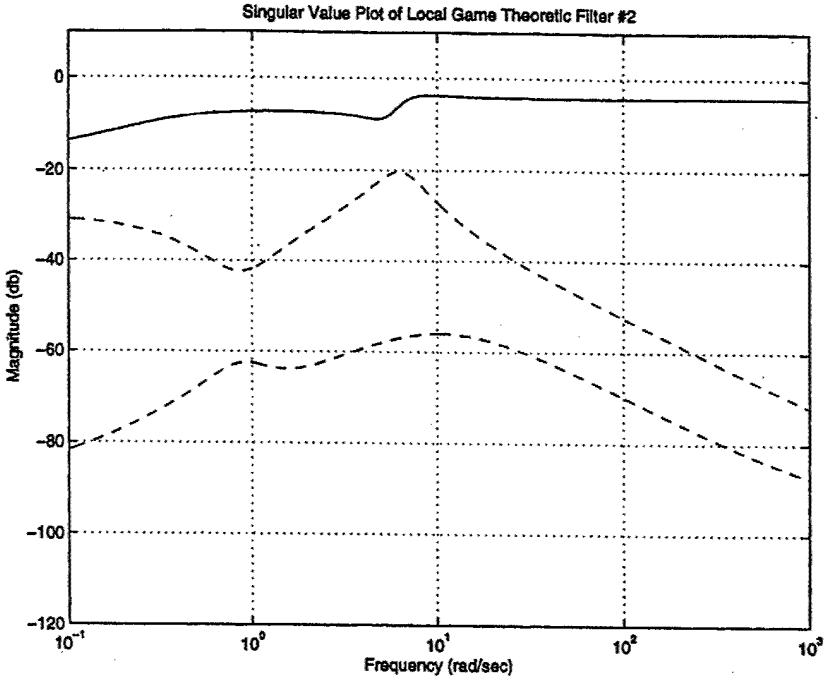


Fig. 3 Platoon example—signal transmission in local detection filter on car # 2

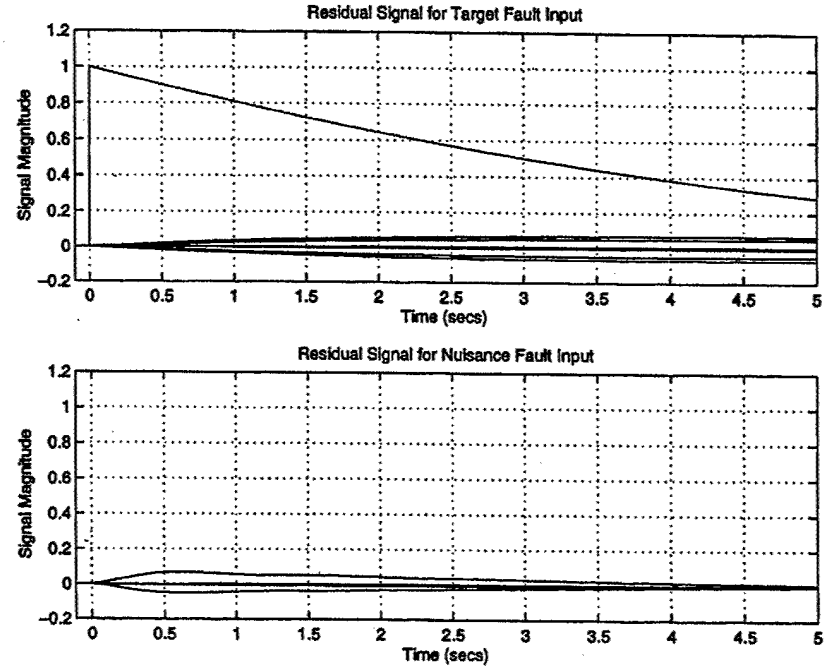


Fig. 4 Platoon example—failure signal response of the decentralized fault detection filter # 1

$$L^1 = 1000\times\begin{bmatrix} -0.0060 & 0.0007 & 0.0112 & 0.1432 & -0.0004 & -0.1021 & -0.0683 \\ -0.0973 & 1.3243 & 0.0971 & -0.0185 & -0.0000 & -0.0205 & 0.1017 \\ 0.0622 & -0.5231 & -0.0386 & 0.0079 & 0.0000 & 0.0099 & -0.0388 \\ 0.5139 & -6.9936 & -0.5124 & 0.0959 & 0.0000 & 0.1058 & -0.5385 \\ -0.0041 & 0.0581 & 0.0057 & -0.0141 & 0.0000 & -0.0118 & -0.0042 \\ 0.0002 & -0.0028 & 0.0151 & -0.0910 & -0.0000 & -0.1169 & -0.0915 \end{bmatrix}$$

for car #1. The transmission properties of the filter are depicted in Fig. 2. The minimum separation over frequency is only about 10 dB, but the filter has particularly good separation in the high frequency range. For car #2, the same weightings, adjusted for the different dimensions of the car #2 dynamics,

$$M^2 = 10 \times I_4, \quad V^2 = \text{diag}[1\ 1\ 10\ 1\ 1\ 1\ 1\ 1], \quad Q^2 = I_8, \quad \gamma_2 = 0.18,$$

leads to

$$L^2 = 1000 \times \begin{bmatrix} -0.0001 & -0.0000 & 0.0000 & -0.0000 & -0.0000 & -0.0003 & -0.0002 & 1.1715 \\ 0.0001 & 0.0000 & -0.0000 & 0.0000 & 0.0000 & 0.0003 & 0.0002 & -1.2154 \\ -0.0000 & -0.0000 & 0.0000 & -0.0000 & -0.0000 & 0.0000 & 0.0001 & -0.0247 \\ -0.0000 & 0.0001 & 0.0000 & -0.0000 & -0.0000 & -0.0001 & -0.0000 & -0.0128 \\ 0.0081 & -0.0010 & -0.0000 & 0.0009 & 0.0000 & 0.0020 & 0.0012 & -0.0002 \\ 0.0027 & 0.0005 & 0.0003 & -0.0542 & 0.0001 & -0.0106 & -0.0146 & -0.0306 \\ -0.0034 & 0.0003 & 0.0005 & 0.0182 & -0.0002 & -0.0482 & -0.0299 & -0.0233 \\ 0.1650 & -2.2230 & -0.0162 & 0.0269 & 0.0000 & 0.0286 & -0.1752 & 0 \end{bmatrix}.$$

The transmission properties for this particular filter are depicted in Fig. 3. The reader should notice the similarities in the level of performance between this filter and the one designed for car #1.

Finally, for the global system, a fault detection filter for range sensor health monitoring in the platoon is found by solving the corresponding Riccati equation with the weightings:

$$\gamma V^{-1} = \text{diag}(1,100,100,1,1,1,1,1,100,100,1,1,1,1,1), \quad Q = I_{15},$$

$$M = 100 \times I_8, \qquad \gamma = 0.18.$$

For the global system, however, we are not interested in finding a gain for a global filter, but in obtaining a global Riccati solution, Π, for use in determining the blending matrices,

$$G^j = \gamma \Pi^{-1} S^j \left(\frac{\alpha_j}{\gamma_j} \Pi^j \right)$$

The connecting matrices, S^j, are taken to be the pseudo-inverses of E^j. As the dimensions of these matrices are quite large,[2] we cannot list them in this paper.

Note that we use our design freedom in V. The reason for this is that if we had not used this freedom and chosen

$$V = \begin{bmatrix} V^1 & 0 \\ 0 & V^2 \end{bmatrix},$$

the response of the filter to the target fault input would have been unsatisfactory. In Fig. 4, we show this response, which is constructed by implementing our decentralized estimator, (10). In this figure, the time history of the failure signal, z, is shown when the system is driven by a fault in the range sensor and a fault in the longitudinal accelerometer. The range sensor is the target fault and as the corresponding plot in Fig. 4 shows, this fault is seen almost immediately in the residual. Better yet, its presence is seen over a sustained period. Had we not adjusted the weightings in V, the time constants in our decentralized filter would have been too small resulting in a target fault response that dies away too quickly.

The reader should note that the responses seen in Fig. 4 can be understood to be the result of the direct feedthrough of the fault into z since a range fault goes directly to the global measurement vector, y. The longitudinal accelerometer is the nuisance fault and we see in the corresponding plot that this failure is also fed through to the residual but at a much smaller magnitude. A reasonably well-designed redundancy management system should, thus, be able to detect the range sensor fault no matter the behavior of the longitudinal accelerometer.

[2] Π is 15×15 for instance

6 Conclusions

In this paper, we have introduced a decentralized fault detection filter that provides an alternative way to monitor large-scale systems for faults. The resulting filter has additional fault tolerance, because it can check the health of its contituent sensors prior to deriving the top level estimate, and it is easily scalable. We have also introduced a logical and theoretically rigorous method for decomposing large, global systems into smaller, local ones using minimum realizations. An example based upon the linearization of a nonlinear car model is given to illustrate our results.

References

[1] White, J. E., and Speyer, J. L., 1987, "Detection Filter Design: Spectral Theory and Algorithms," IEEE Trans. Autom. Control, **AC-32**, pp. 593–603.

[2] Massoumnia, M.-A., 1986, "A Geometric Approach to the Synthesis of Failure Detection Filters," IEEE Trans. Autom. Control, **AC-31**, pp. 839–846.

[3] Ding, X., and Frank, P. M., 1989, "Fault Detection via Optimally Robust Detection Filters," *Proceedings of the 28th Conference on Decision and Control*, Tampa, FL, pp. 1767–1772, Institute of Electrical and Electronic Engineers.

[4] Chung, W. H., and Speyer, J. L., 1998, "A Game Theoretic Fault Detection Filter," IEEE Trans. Autom. Control, **AC-43**, pp. 145–161.

[5] Patton, R., and Chen, J., 1992, "Robust Fault Detection of Jet Engine Sensor Systems Using Eigenstructure Assignment," J. Guid. Control Dyn., **15**, pp. 1491–1497.

[6] Patton, R., Chen, J., and Millar, J., 1991, "A Robust Disturbance Decoupling Approach to Fault Detection in Process Systems," *Proceedings of the 30th Conference on Decision and Control*, Brighton, UK, pp. 1543–1548, Institute of Electrical and Electronic Engineers.

[7] Speyer, J. L., 1979, "Computation and Transmission Requirements for a Decentralized Linear-Quadratic-Gaussian Control Problem," IEEE Trans. Autom. Control, **AC-24**, pp. 266–269.

[8] Willsky, A. S., Bello, M. G., Castanon, D. A., Levy, B. C., and Verghese, G. C., 1982, "Combining and Updating of Local Estimates and Regional Maps Along Sets of One-Dimensional Tracks," IEEE Trans. Autom. Control, **AC-27**, pp. 799–813.

[9] Kerr, T., 1981, "Decentralized Filtering and Redundancy Management for Multisensor Integrated Navigation Systems," IEEE Transactions on Aerospace and Electronic Systems, **AES-23(1)**, pp. 83–119.

[10] Chung, W. H., and Speyer, J. L., 1995, "A General Framework for Decentralized Estimation," *Proceedings of the 1995 American Control Conference*, Seattle, WA, American Control Council.

[11] Chung, W. H., 1997, "Game Theoretic and Decentralized Estimation for Fault Detection," PhD thesis, University of California, Los Angeles.

[12] Jang, J., and Speyer, J. L., 1994, "Decentralized Game-Theoretic Filters," *Proceedings of the 1994 American Control Conference*, Baltimore, MD, pp. 3379–3384, American Control Council.

[13] Kalman, R., 1964, "When is a Linear Control System Optimal?," ASME J. Basic Eng., **86**, pp. 51–50.

[14] Beard, R. V., 1971, "Failure Accommodation in Linear Systems through Self-Reorganization," PhD thesis, Massachusetts Institute of Technology.
[15] Douglas, R. K., 1993, "Robust Detection Filter Design," PhD thesis, University of Texas at Austin.
[16] Massoumnia, M.-A., Verghese, G. C., and Willsky, A. S., 1989, "Fault Detection and Identification," IEEE Trans. Autom. Control, **AC-34**, pp. 316–321.
[17] Green, M., and Limebeer, D. J., 1995, *Linear Robust Control*, Prentice-Hall, NJ.
[18] Douglas, R. K., Speyer, J. L., Mingori, D. L., Chen, R. H., Malladi, D. P., and Chung, W. H., 1995, "Fault Detection and Identification with Application to Advanced Vehicle Control Systems," Research Report UCB-ITS-PRR-95-26, California PATH.
[19] Douglas, R. K., Speyer, J. L., Mingori, D. L., Chen, R. H., Malladi, D. P., and Chung, W. H., 1996, "Fault Detection and Identification with Application to Advanced Vehicle Control: Final Report," Research Report UCB-ITS-PRR-96-25, California PATH.

Appendix J

"Target Association Using Detection Methods,"

Jonathan D. Wolfe and Jason L. Speyer,

AIAA J. Guidance, Control, and Dynamics, **Vol. 25, No. 6, November-December, 2002.**

JOURNAL OF GUIDANCE, CONTROL, AND DYNAMICS
Vol. 25, No. 6, November–December 2002

Target Association Using Detection Methods

Jonathan D. Wolfe* and Jason L. Speyer†
University of California, Los Angeles, Los Angeles, California 90095-1597

A residual-based scheme is presented for solving the radar track-to-track association problem using bearings-only measurements. To accomplish track association between two stations, the residuals of a bank of nonlinear filters called modified gain extended Kalman filters are analyzed. Once tracks have been associated between two stations, tracks from additional stations may be associated with tracks from the first two stations by checking algebraic parity equations. Traditional track association methods rely on the local stations' estimated target positions and error variances. These local estimates may be quite inaccurate or even divergent when using bearings-only measurements. Our method bypasses this difficulty because our filters use raw data from multiple stations. An example demonstrates that our methods yield results superior to those of standard methods.

I. Introduction

SUPPOSE that several spatially distributed radar installations are each tracking several targets. Associating a given target to its track at each of the radar stations is an important issue, which the radar literature refers to as the track-to-track association problem. Suppose further that the stations use passive sensors that only measure bearings to the target, without measuring range. In this paper, we outline a strategy for solving this association problem by analyzing measurement residuals.

Bearings-only observation functions fall into two special classes of nonlinear functions, called modifiable and approximately modifiable nonlinearities, which are defined as follows:

Definition 1. A time-varying function $f:\mathbb{R}^n \to \mathbb{R}^q$ is called modifiable if there exists an operator $A:\mathbb{R}^q \times \mathbb{R}^n \to \mathbb{R}^{q\times n}$ such that for any $x, \bar{x} \in \mathbb{R}^n$,

$$f(x) - f(\bar{x}) = A[f(x), \bar{x}](x - \bar{x}) \tag{1}$$

Definition 2. A time-varying function $f:\mathbb{R}^n \to \mathbb{R}^q$ is called approximately modifiable if there exists a region $\mathcal{D} \subset \mathbb{R}^n$ and operators $A:\mathbb{R}^q \times \mathbb{R}^n \to \mathbb{R}^{q\times n}$ and $\mathcal{E}:\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{n\times n}$ such that for any x, $\bar{x} \in \mathcal{D}$,

$$f(x) - f(\bar{x}) = [A(f(x), \bar{x}) + \mathcal{E}(x, x - \bar{x})](x - \bar{x}) \tag{2}$$

where $\lim_{\|x-\bar{x}\|\to 0} \|\mathcal{E}(x, x-\bar{x})\|/\|A(f(x), \bar{x})\| = 0$.

Song and Speyer's modified gain extended Kalman filter (MGEKF)[1] is a globally convergent, unbiased, nonlinear observer for systems whose measurement functions are modifiable or approximately modifiable. In this paper, the observers we design for bearings-only track association are MGEKFs.

An early attempt at solving the track-to-track association problem was made by Singer and Kanyuck.[2] In their paper, they incorrectly assumed that estimation errors local to each station were uncorrelated. Bar-Shalom,[3] Bar-Shalom and Fortmann,[4] and Bar-Shalom and Campo[5] later corrected this error by accounting for the correlation between the local estimation errors due to the common process noise of the target. Later researchers have integrated the problem of track association directly into the process of separating the measurements corresponding to actual targets from clutter.[6,7] In all of these references, it is assumed that both range and bearings were measured. In some of these references, the possibility of using a MGEKF to handle the situation of bearings-only measurements is mentioned, but none have a discussion of the details of such an implementation, in particular problems associated with the asymmetry of single station estimation errors. Estimates based on bearings-only measurements from a single station are especially uncertain along the line between the target and the receiver. This uncertainty is reduced when measurements from physically separated stations are used. Our method attempts to take advantage of this phenomenon by using estimates constructed from several stations' measurements.

The paper is organized as follows. We show in Sec. II that bearings-only measurement functions are modifiable. (Prior results only showed that they were approximately modifiable.[1]) We then demonstrate in Sec. III that incorrect associations between two radar stations can be interpreted as sensor faults, so that a bank of modified-gain fault detection filters can be used to determine the track associations. Section IV contains the main result, an algorithm for solving the bearings-only track association problem. The application of this algorithm to an example in Sec. V compares our approach to a conventional track association method. Section VI concludes the paper.

In the sequel, inertial Cartesian coordinates describe the motion of each target in three dimensions via the state vector

$$\boldsymbol{x}^t = [X^t \quad Y^t \quad Z^t \quad \dot{X}^t \quad \dot{Y}^t \quad \dot{Z}^t \quad \ddot{X}^t \quad \ddot{Y}^t \quad \ddot{Z}^t]^T \tag{3}$$

and the dynamics of each target are assumed to be of the form

$$\boldsymbol{x}^t(k+1) = \boldsymbol{A}(k)\boldsymbol{x}^t(k) + \boldsymbol{B}(k)\boldsymbol{w}^t(k) \tag{4}$$

Note that we include an acceleration state to model maneuvering target dynamics.

II. Modifiability of Bearings-Only Measurements

Song and Speyer[1] showed that the azimuth angle $az_s^t \in [-\pi/2, \pi/2)$ and the elevation angle $el_s^t \in [-\pi/2, \pi/2)$ from station s to target t, as shown in Fig. 1, are modifiable and approximately modifiable, respectively. The region $\mathcal{D}$ in which the elevation angle was approximately modifiable excluded an ellipsoidal region near the sensor, making their algorithms difficult to implement for situations where the angular sensor gets close to the target, for example, in the terminal guidance of a missile. We improve this situation somewhat by introducing the new angle $\Psi_s^t \in [-\pi/2, \pi/2)$ and describing the position of the target in terms of Ψ_s^t and $\Phi_s^t \triangleq az_s^t$. Note that Ψ_s^t can be calculated from az_s^t and el_s^t via the equation

$$\Psi_s^t = \tan^{-1}\left(\frac{Z_s^t}{X_s^t}\right) = \tan^{-1}\left(\frac{\tan el_s^t}{\cos az_s^t}\right) \tag{5}$$

This section is devoted to proving that the measurement function for Ψ_s^t is modifiable.

Let $\bar{\boldsymbol{x}}^t$ be an estimate of $\boldsymbol{x}^t$ and assume that the position of the measurement station in inertial space, $\boldsymbol{x}_s = [X_s \quad Y_s \quad Z_s]$, is known.

Received 4 October 2001; revision received 1 May 2002; accepted for publication 12 June 2002.

*Research Engineer, Department of Mechanical and Aerospace Engineering.
†Professor, Department of Mechanical and Aerospace Engineering. Fellow AIAA.

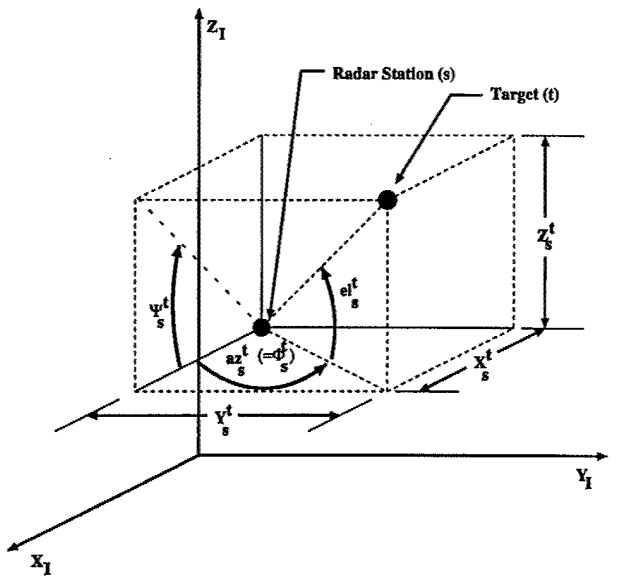


Fig. 1 Angles for target bearings.

Then X_s^t, Y_s^t, Z_s^t, $\bar{X}_s^t$, $\bar{Y}_s^t$, and $\bar{Z}_s^t$ can be computed by taking the difference between elements of x^t, $\bar{x}^t$, and x_s.

Suppose that station s measures the bearings of target t with the measurement vector z_s^t. Define $h_s(x^t)$ by

$$h_s(x^t) \triangleq \begin{bmatrix} \Phi_s^t \\ \Psi_s^t \end{bmatrix} = z_s^t \tag{6}$$

The measurement residual corresponding to $h_s(x^t)$ is then

$$h_s(x^t) - h_s(\bar{x}^t) = \begin{bmatrix} \tan^{-1}\left(Y_s^t/X_s^t\right) - \tan^{-1}\left(\bar{Y}_s^t/\bar{X}_s^t\right) \\ \tan^{-1}\left(Z_s^t/X_s^t\right) - \tan^{-1}\left(\bar{Z}_s^t/\bar{X}_s^t\right) \end{bmatrix} \triangleq \begin{bmatrix} \tan^{-1}\alpha \\ \tan^{-1}\beta \end{bmatrix} \tag{7}$$

Applying the trigonometric identity

$$\tan^{-1}(a) - \tan^{-1}(b) = \tan^{-1}[(a-b)/(1+ab)]$$

we obtain

$$\begin{bmatrix} \tan^{-1}\alpha \\ \tan^{-1}\beta \end{bmatrix} = \begin{bmatrix} \tan^{-1}\dfrac{\left(Y_s^t/X_s^t\right) - \left(\bar{Y}_s^t/\bar{X}_s^t\right)}{1 + \left(Y_s^t/X_s^t\right)\left(\bar{Y}_s^t/\bar{X}_s^t\right)} \\ \tan^{-1}\dfrac{\left(Z_s^t/X_s^t\right) - \left(\bar{Z}_s^t/\bar{X}_s^t\right)}{1 + \left(Z_s^t/X_s^t\right)\left(\bar{Z}_s^t/\bar{X}_s^t\right)} \end{bmatrix} \tag{8}$$

$$\begin{bmatrix} \tan^{-1}\alpha \\ \tan^{-1}\beta \end{bmatrix} = \begin{bmatrix} \tan^{-1}\left(\dfrac{Y_s^t\bar{X}_s^t - \bar{Y}_s^t X_s^t}{X_s^t\bar{X}_s^t + Y_s^t\bar{Y}_s^t}\right) \\ \tan^{-1}\left(\dfrac{Z_s^t\bar{X}_s^t - \bar{Z}_s^t X_s^t}{X_s^t\bar{X}_s^t + Z_s^t\bar{Z}_s^t}\right) \end{bmatrix} \tag{9}$$

Define

$$H\left(z_s^t\right) \triangleq \begin{bmatrix} \sin\left(\Phi_s^t\right) & -\cos\left(\Phi_s^t\right) & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \sin\left(\Psi_s^t\right) & 0 & -\cos\left(\Psi_s^t\right) & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{10}$$

Let $d_1 \triangleq \sqrt{[(X_s^t)^2 + (Y_s^t)^2]}$, $D_1 \triangleq d_1/[X_s^t\bar{X}_s^t + Y_s^t\bar{Y}_s^t]$, $d_2 \triangleq \sqrt{[(X_s^t)^2 + (Z_s^t)^2]}$, and $D_2 \triangleq d_2/[X_s^t\bar{X}_s^t + Z_s^t\bar{Z}_s^t]$. Note also that $\sin(\Phi_s^t) = Y_s^t/d_1$, $\cos(\Phi_s^t) = X_s^t/d_1$, $\sin(\Psi_s^t) = Z_s^t/d_2$, and $\cos(\Psi_s^t) = X_s^t/d_2$. Therefore, we can express D_1 and D_2 as functions of the estimates and measured angles:

$$D_1 = D_1\left(z_s^t, \bar{x}^t\right) = 1/\left[\cos\left(\Phi_s^t\right)\bar{X}_s^t + \sin\left(\Phi_s^t\right)\bar{Y}_s^t\right]$$

$$D_2 = D_2\left(z_s^t, \bar{x}^t\right) = 1/\left[\cos\left(\Psi_s^t\right)\bar{X}_s^t + \sin\left(\Psi_s^t\right)\bar{Z}_s^t\right]$$

If we express the trigonometric functions in $H(z_s^t)$, D_1, and D_2 in terms of X_s^t, Y_s^t, Z_s^t, $\bar{X}_s^t$, $\bar{Y}_s^t$, and $\bar{Z}_s^t$, we can write Eq. (9) as a function of z_s^t and $\bar{x}^t$:

$$\begin{bmatrix} \alpha\left(z_s^t, \bar{x}^t\right) \\ \beta\left(z_s^t, \bar{x}^t\right) \end{bmatrix} = \begin{bmatrix} D_1\left(z_s^t, \bar{x}^t\right) & 0 \\ 0 & D_2\left(z_s^t, \bar{x}^t\right) \end{bmatrix} H\left(z_s^t\right)\left[\bar{x}^t - x_s\right] \tag{11}$$

Finally, we can rewrite Eq. (11) as

$$\begin{bmatrix} -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1/\alpha\left(z_s^t, \bar{x}^t\right) & 0 \\ 0 & 1/\beta\left(z_s^t, \bar{x}^t\right) \end{bmatrix} \begin{bmatrix} D_1\left(z_s^t, \bar{x}^t\right) & 0 \\ 0 & D_2\left(z_s^t, \bar{x}^t\right) \end{bmatrix} \times H\left(z_s^t\right)\left[x^t - x^t - \bar{x}^t + x_s\right] \tag{12}$$

and combine it with Eq. (7) to obtain $h_s(x^t)$ in modifiable form,

$$h_s(x^t) - h_s(\bar{x}^t) = \begin{bmatrix} -\dfrac{D_1\left(z_s^t, \bar{x}^t\right)\tan^{-1}\alpha\left(z_s^t, \bar{x}^t\right)}{\alpha\left(z_s^t, \bar{x}^t\right)} & 0 \\ 0 & -\dfrac{D_2\left(z_s^t, \bar{x}^t\right)\tan^{-1}\beta\left(z_s^t, \bar{x}^t\right)}{\beta\left(z_s^t, \bar{x}^t\right)} \end{bmatrix} \times H\left(z_s^t\right)\left[x^t - \bar{x}^t\right] \tag{13}$$

where we have made use of the identity

$$H\left(z_s^t\right)\left[x_s - x^t\right] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Thus, we have replaced the elevation angle el_s^t, from which Song and Speyer[1] produced an approximately modifiable function with a new angle Ψ_s^t. Like the azimuth angle Φ_s^t, angle Ψ_s^t leads to modifiable measurement functions.

III. Converting Incorrect Associations into Sensor Faults

Suppose that station s can view several targets, indexed by i, and measures the bearings of each target. Then each of these measurements z_s^i is generated by $h_s(x^i)$, as in Eq. (6). Now suppose that another station, using its local observations, generates a state estimate of one of the targets that station s views. This estimate $\bar{x}^j$ corresponds to x^j, the true state of the jth target at station s, but neither station knows the value of index j. Our goal is to determine which of the tracks at station s is the jth one, using only $\{z_s^i\}$, the measurements local to station s, and $\bar{x}^j$, the other station's state estimate of one of the targets.

To this end, let us form the following error residual between the estimate $\bar{x}^j$ and the measurement z_s^i, making use of the result from the preceding section:

$$z_s^i - h_s(\bar{x}^j) = h_s(x^i) - h_s(\bar{x}^j) = G\left(z_s^i, \bar{x}^j\right)(x^i - \bar{x}^j) \tag{14}$$

where from Eq. (13)

$$G\left(z_s^i, \bar{x}^j\right) = \begin{bmatrix} -\dfrac{D_1\left(z_s^i, \bar{x}^j\right)\tan^{-1}\alpha\left(z_s^i, \bar{x}^j\right)}{\alpha\left(z_s^i, \bar{x}^j\right)} & 0 \\ 0 & -\dfrac{D_2\left(z_s^i, \bar{x}^j\right)\tan^{-1}\beta\left(z_s^i, \bar{x}^j\right)}{\beta\left(z_s^i, \bar{x}^j\right)} \end{bmatrix} \times H\left(z_s^i\right) \tag{15}$$

By introducing a zero term into the measurement residual, we can rephrase it as

$$z_s^i - h_s(\bar{x}^j) = h_s(x^i) - h_s(\bar{x}^j) \tag{16}$$

$$z_s^i - h_s(\bar{x}^j) = G\left(z_s^i, \bar{x}^j\right)(x^i - \bar{x}^j) \tag{17}$$

$$z_s^i - h_s(\bar{x}^j) = G\left(z_s^i, \bar{x}^j\right)(x^i - \bar{x}^j + x^j - x^j) \tag{18}$$

$$z_s^i - h_s(\bar{x}^j) = G\left(z_s^i, \bar{x}^j\right)(x^j - \bar{x}^j) + G\left(z_s^i, \bar{x}^j\right)(x^i - x^j) \tag{19}$$

$$z_s^i - h_s(\bar{x}^j) = G\left(z_s^i, \bar{x}^j\right)(x^j - \bar{x}^j) + \mu_s^{ij} \tag{20}$$

where $\mu_s^{ij} \triangleq G(z_s^i, \bar{x}^j)(x^i - x^j)$ represents the difference between x^i and x^j as a sensor fault. If $i = j$, we have correctly guessed the association between measurement and estimate, and there is no fault ($\mu_s^{ij} = 0$). If $i \neq j$, then $\mu_s^{ij} \neq 0$, playing the role of a sensor fault in the residual.

IV. Algorithm for Track Association from Bearings-Only Measurements via Fault Detection Filters

Suppose that there are S radar stations, with known inertial coordinates, that make bearings-only measurements in three-space of T different targets. We assume that all measurements at each station have been grouped as tracks of each target visible at that station using conventional means.[4,8,9] In this section, we propose an algorithm for associating the tracks at all stations to their corresponding targets.

Assume that each measurement station s is located at known inertial coordinates (X_s, Y_s, Z_s). Let $\bar{x}^{1i}$ denote a fault detection filter's estimate of the target corresponding to the ith track at the first station. The bearings-only measurement function for the station s of the same target is thus

$$h_s(x^{1i}) \triangleq \begin{bmatrix} \tan^{-1}\left(\dfrac{Y^{1i} - Y_s}{X^{1i} - X_s}\right) \\ \tan^{-1}\left(\dfrac{Z^{1i} - Z_s}{X^{1i} - X_s}\right) \end{bmatrix}$$

From the results of the preceding section, the error residual of track j at any station s, generated by target i at station 1, is given by

$$z_s^j - h_s(\bar{x}^{1i}) \approx G\left(z_s^j, \bar{x}^{1i}\right)(x^{1i} - \bar{x}^{1i}) + \mu_s^{ij} + v^j \tag{21}$$

where $G(z_s^j, \bar{x}^{1i})$ is given by Eq. 15 and the sensor noise is

$$v^j = \mathcal{N}(0, V^j)$$

The approximate structure of Eq. (21) is due to the replacement of the measurement function in $G(\cdot, \cdot)$ with the actual measurement (see Song and Speyer[1]). Note that, by default, $\mu_1^{ii} = 0, \forall i = 1, \ldots, T$.

The following algorithm, illustrated in Fig. 2, associates tracks between stations.

Algorithm (track association):

1) Let $i = 1$.

2) Run a bank of T detection filters that operate on data from stations 1 and 2, where the jth filter attempts to detect μ_2^{ij}. Each filter is constructed using the dynamic detection filter procedure given next. All but one of these detection filters should register a fault. The track corresponding to the filter that detected no fault is associated with z_1^i. Without loss of generality, label this track z_2^i.

3) For each track z_s^l, $s = 3, \ldots, S$, $l = 1, \ldots, T$, perform the algebraic parity test given subsequently. If the result of the parity test is zero, then z_s^l is associated with z_1^i and z_2^i.

4) If $i < T$, increment i by 1 and go to step 2. If $i = T$, we have completed the track association procedure.

Note that estimates obtained in step 2 are used in step 3. Therefore, stations 1 and 2 should be chosen to maximize observability of the targets.

Dynamic Detection Filter

For any estimator of x^{1i}, the estimation residual determined by the measurements z_1^i and z_2^j will not converge to values near zero unless z_1^i and z_2^j correspond to the same target. One such estimator is the MGEKF[1] given as

$$\bar{x}^{1i}(k+1) = A(k)\hat{x}^{1i}(k) \tag{22}$$

$$r^{ij}(k) = \begin{Bmatrix} z_1^i(k) - h_1[\hat{x}^{1i}(k)] \\ z_2^j(k) - h_2[\hat{x}^{1i}(k)] \end{Bmatrix} \tag{23}$$

$$\hat{x}^{1i}(k) = \bar{x}^{1i}(k) + K^{ij}(k)r^{ij}(k) \tag{24}$$

$$M^{ij}(k+1) = A(k)P^{ij}(k)A^T(k) + Q(k) \tag{25}$$

$$\bar{h}_{x^{1i}(k)} = \begin{bmatrix} -\dfrac{\bar{Y}^{1i} - Y_1}{\left\{1 + \left[\left(\bar{Y}^{1i} - Y_1\right)/\left(\bar{X}^{1i} - X_1\right)\right]^2\right\}\left(\bar{X}^{1i} - X_1\right)^2} & \dfrac{1}{\left\{1 + \left[\left(\bar{Y}^{1i} - Y_1\right)/\left(\bar{X}^{1i} - X_1\right)\right]^2\right\}\left(\bar{X}^{1i} - X_1\right)^2} & \cdots \\ -\dfrac{\bar{Z}^{ij} - Z_1}{\left\{1 + \left[\left(\bar{Z}^{ij} - Z_1\right)/\left(\bar{X}^{1i} - X_1\right)\right]^2\right\}\left(\bar{X}^{1i} - X_1\right)^2} & 0 \\ -\dfrac{\bar{Y}^{1i} - Y_2}{\left\{1 + \left[\left(\bar{Y}^{1i} - Y_2\right)/\left(\bar{X}^{1i} - X_2\right)\right]^2\right\}\left(\bar{X}^{1i} - X_2\right)^2} & \dfrac{1}{\left\{1 + \left[\left(\bar{Y}^{1i} - Y_2\right)/\left(\bar{X}^{1i} - X_2\right)\right]^2\right\}\left(\bar{X}^{1i} - X_2\right)^2} & \cdots \\ -\dfrac{\bar{Z}^{ij} - Z_2}{\left\{1 + \left[\left(\bar{Z}^{ij} - Z_2\right)/\left(\bar{X}^{1i} - X_2\right)\right]^2\right\}\left(\bar{X}^{1i} - X_2\right)^2} & 0 \end{bmatrix}$$

$$\cdots \begin{bmatrix} 0 & 0\ 0\ 0\ 0\ 0\ 0 \\ \dfrac{1}{\left\{1 + \left[\left(\bar{Z}^{ij} - Z_1\right)/\left(\bar{X}^{1i} - X_1\right)\right]^2\right\}\left(\bar{X}^{1i} - X_1\right)^2} & 0\ 0\ 0\ 0\ 0\ 0 \\ 0 & 0\ 0\ 0\ 0\ 0\ 0 \\ \dfrac{1}{\left\{1 + \left[\left(\bar{Z}^{ij} - Z_2\right)/\left(\bar{X}^{1i} - X_2\right)\right]^2\right\}\left(\bar{X}^{1i} - X_2\right)^2} & 0\ 0\ 0\ 0\ 0\ 0 \end{bmatrix} \tag{26}$$

Table 1 Radar station positions

Station identification	x Position, m	y Position, m	z Position, m
1	50	1	0
2	50,000	1	50
3	25,000	−400	100

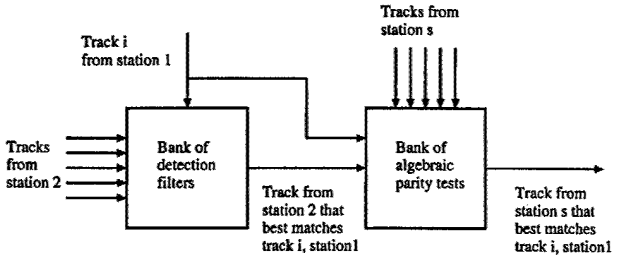


Fig. 2 Track association procedure.

$$K^{ij}(k) = M^{ij}(k)\bar{h}^T_{x^{1i}(k)}\left[\bar{h}_{x^{1i}(k)}M^{ij}(k)\bar{h}^T_{x^{1i}(k)} + V^{ij}(k)\right]^{-1} \tag{27}$$

$$\tilde{G}\left[z_1^i(k), z_2^j(k), \bar{x}^{1i}(k)\right] = \begin{bmatrix} G\left(z_1^i(k), \bar{x}^{1i}(k)\right) \\ G\left(z_2^j(k), \bar{x}^{1i}(k)\right) \end{bmatrix} \tag{28}$$

$$P^{ij}(k) = \left\{I - K^{ij}(k)\tilde{G}\left[z_1^i(k), z_2^j(k), \bar{x}^{1i}(k)\right]\right\}M^{ij}(k)$$
$$\times \left\{I - K^{ij}(k)\tilde{G}\left[z_1^i(k), z_2^j(k), \bar{x}^{1i}(k)\right]\right\}^T$$
$$+ K^{ij}(k)(V^{ij})^{-1}(k)(K^{ij})^T(k) \tag{29}$$

where

$$V^{ij}(k) = \text{diag}\{V^i, V^j\} \tag{30}$$

The weighted innovations process of the MGEKF,

$$\nu^{ij}(k) = \left[\bar{h}_{x^{1i}(k)}M^{ij}(k)\bar{h}^T_{x^{1i}(k)} + V^{ij}(k)\right]^{-\frac{1}{2}} r^{ij}(k) \tag{31}$$

should be close to a zero-mean, unit variance white noise sequence only if z_1^i and z_2^j correspond to the same target.

Algebraic Parity Test

This test determines if z_s^l, $S \geq s > 2$, $T \geq l \geq 1$, is associated with z_1^i and z_2^j, where z_1^i and z_2^j are already known to be associated with each other. Suppose that $\hat{x}^{1i}$ is the state estimate generated by z_1^i and z_2^j. Then, if z_s^l is associated with the tracks z_1^i and z_2^j,

$$\nu(k)_l^{1i} \triangleq \left[\bar{h}_{x^{1i}(k)}M^{ij}(k)\bar{h}^T_{x^{1i}(k)} + V^{ij}(k)\right]^{-\frac{1}{2}}\left\{z_s^l(k) - h_s[\hat{x}^{1i}(k)]\right\} \tag{32}$$

should be close to a zero mean, unit variance white noise sequence. Here, the approximate measurement matrix $\bar{h}_{x^{1i}(k)}$ is computed in a manner similar to the first two rows of the matrix in Eq. (26), but referenced to (X_s, Y_s, Z_s), the location of station s, instead of the location of the first station (X_1, Y_1, Z_1). The algebraic parity test is simply to evaluate the parity equation (32).

V. Example

The track association algorithm presented in the last section is applied to simulation data in this section. Three radar installations were located at the positions given by Table 1, and two targets were both modeled as ninth-order linear time-invariant discrete-time systems with the dynamics

$$\dot{x}(t) = Fx(t) + \Gamma w(t) \tag{33}$$

where

$$F \triangleq \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\alpha & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\alpha & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\alpha \end{bmatrix}$$

$$\Gamma \triangleq \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{34}$$

and where w is a zero mean Brownian motion process with covariance $I_{3\times 3}$ and $\alpha = \frac{1}{10}$ is the time constant for the first-order filters that model target maneuvers as colored noise processes. We sample this model at intervals of $T = 0.1$ s to generate the discrete time dynamics

$$x(k+1) = Ax(k) + Bw(k) \tag{35}$$

where

$$A = e^{FT}, \qquad B = \int_0^T e^{Ft}B\,dt, \qquad E[w(k)] = 0_{3\times 1}$$
$$E[w(k)w^T(l)] = I_{3\times 3}\delta_{kl} \tag{36}$$

The targets began the simulation with the initial conditions

$$x_1(0) = [50 \quad 220{,}000 \quad 30{,}000 \quad 250 \quad -1000 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

$$x_2(0) = [50{,}000 \quad 20{,}000 \quad 35{,}000 \quad -250 \quad 1000 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

This configuration corresponds to the two targets initially moving directly toward each other, in a line that almost passes through station 2. In the simulation, they pass closest to each other at $t = 99.2$ s. Each measurement station measures the angles Φ_s^t and Ψ_s^t to each target at every sample time. These measurements are subject to additive, normally distributed zero-mean white measurement noise with standard deviation 1 deg. We assume that the measurement noise is independent between sensors at all stations. Each MGEKF begins with the a priori information

$$\hat{x}^{1i}(0) = [25{,}000 \quad 120{,}000 \quad 32{,}500 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

$$P^{ij}(0) = 10^7 \times I_{9\times 9}$$

Finally, we assume that the local stations were able to separate their measurements from clutter perfectly using methods like those of Reid[9] or Bar-Shalom and Fortmann,[4] or Fortmann and Bar-Shalom.[8]

Figure 3 plots the weighted innovations of a MGEKF that uses measurements from stations 1 and 2 that correspond to the second target, whereas Fig. 4 plots the weighted innovations of a MGEKF that uses measurements that are mismatched. Note that the innovations for the correct match appear to be a zero mean white noise sequence, whereas the innovations for the incorrect match are larger and are not white. To better observe the behavior of these sequences, their means were estimated using a Kalman filter (assuming that each element of the weighted innovation of the MGEKF was a measurement of a process that had integrator dynamics, process noise with

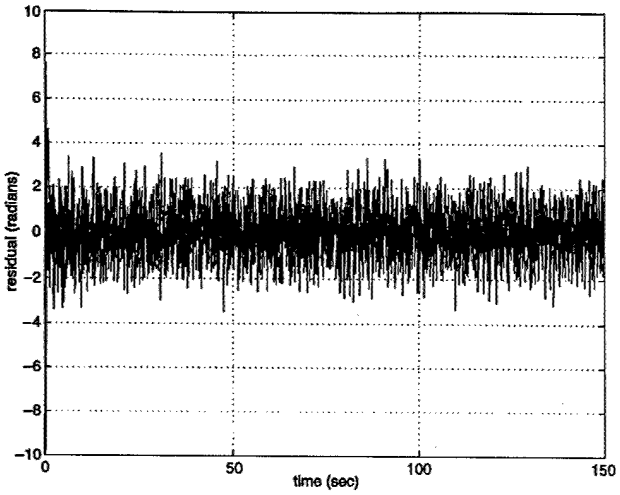


Fig. 3 MGEKF residual for matching tracks.

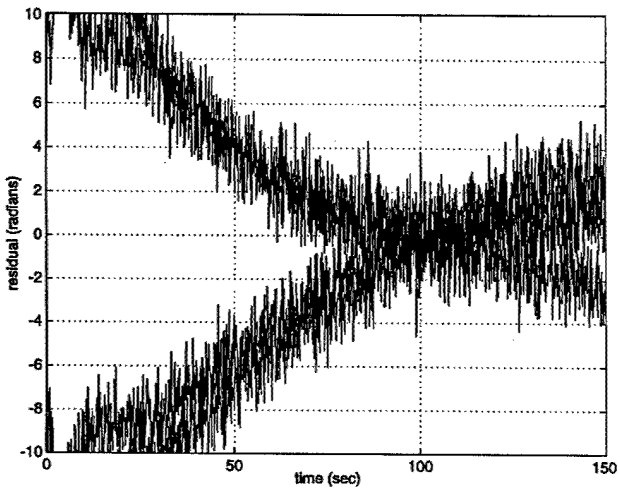


Fig. 4 MGEKF residual for mismatched tracks.

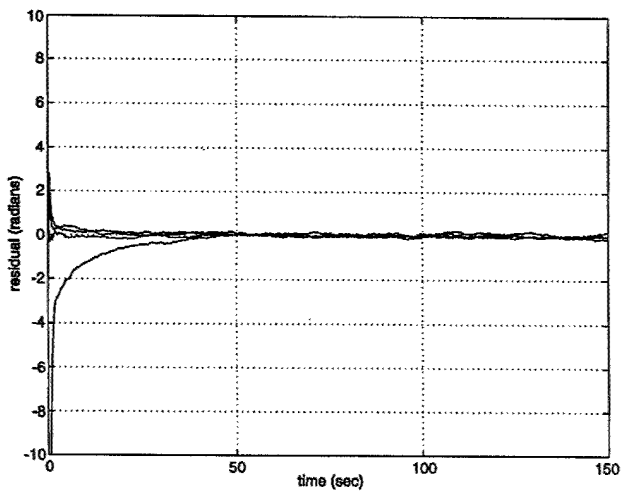


Fig. 5 Filtered MGEKF residual for matching tracks.

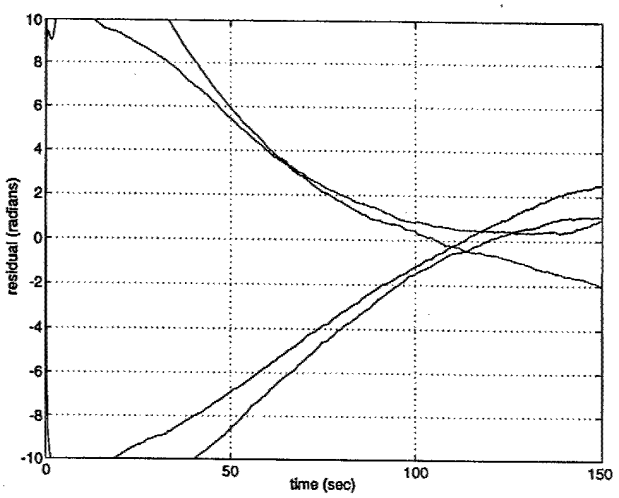


Fig. 6 Filtered MGEKF residual for mismatched tracks.

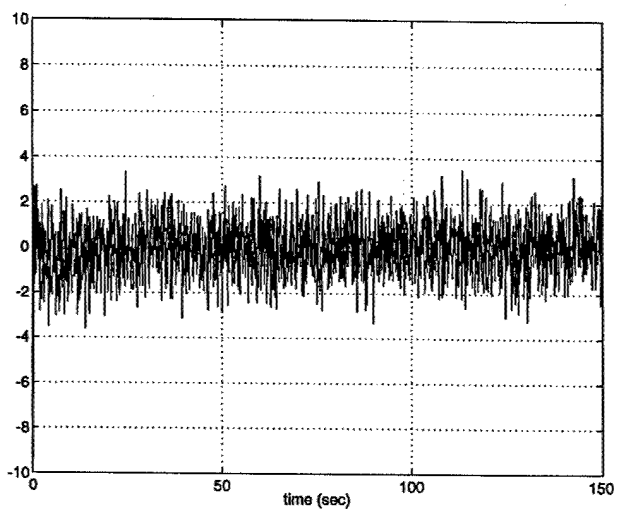


Fig. 7 Parity test residual for matching tracks.

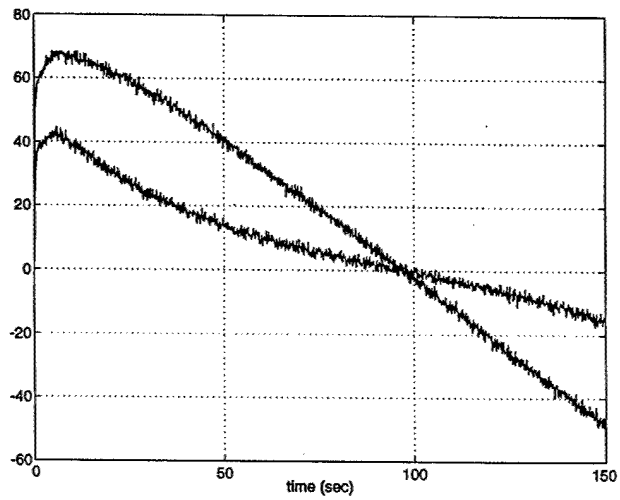


Fig. 8 Parity test residual for mismatched tracks.

covariance 10^{-3}, and measurement noise with covariance 1). These estimates (Figs. 5 and 6) clearly show that the mean corresponding to a mismatch looks nothing like that of the matched case.

After the tracks had been associated between the first two stations, algebraic parity tests attempted to associate the targets observed by the third station relative to those observed by the first and second stations. Two plots of residuals generated by the algebraic parity tests appear in Figs. 7 and 8. Again, the residuals for the mismatch are much larger than those corresponding to a correct association.

For purposes of comparison, Fig. 9 plots the error statistic developed by Bar-Shalom[3] and Bar-Shalom and Fortmann[4] for both a correct and an incorrect track association (using the same data sequences that were used by the filters in Figs. 3 and 4). Note that the chi-squared error statistic does not change much between the matched and mismatched cases. We also noticed that there

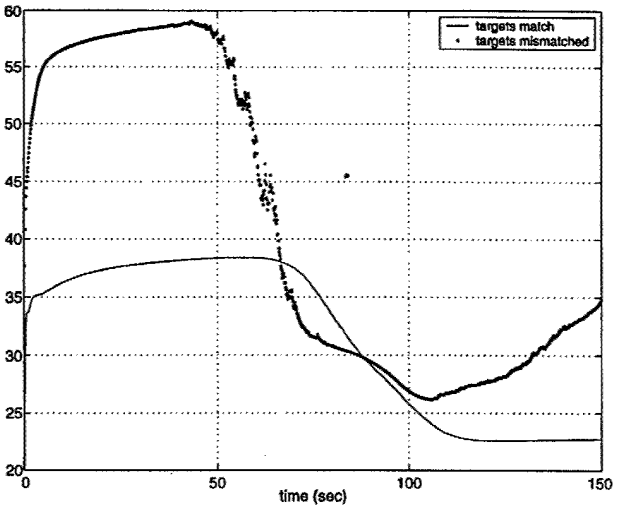


Fig. 9 Error statistic suggested by Bar-Shalom[3] and Bar-Shalom and Fortmann[4]: $(\hat{x}_1 - \hat{x}_2)^T E[(\hat{x}_1 - \hat{x}_2)(\hat{x}_1 - \hat{x}_2)^T](\hat{x}_1 - \hat{x}_2)$.

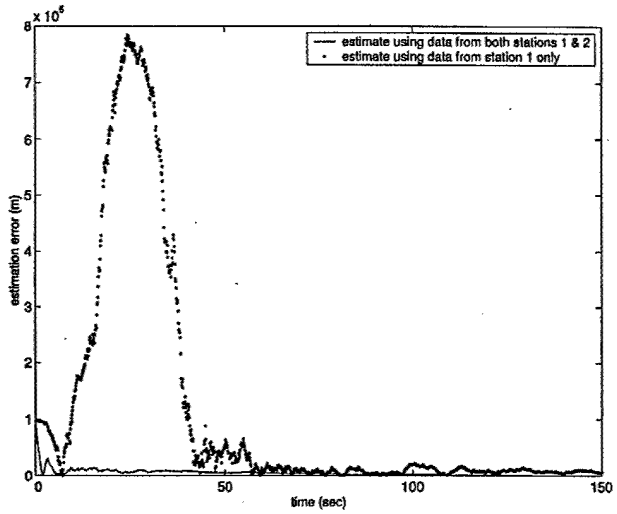


Fig. 10 Euclidean norm of error in tracking target 1.

were several instances where nearly singular matrices were inverted in the algorithm that computes the covariance of the difference between two local estimates.

Part of the reason for this difficulty is explained in Fig. 10, a plot of the Euclidean norm of the estimation error. The solid line corresponds to a MGEKF that uses measurements from both station 1 and 2, whereas the dotted line is from a filter that only used station 1 measurements. Any method that relies on estimates that only use a single station's measurements is subject to a large error. This is not a huge concern for linear estimators, but the matrix $\boldsymbol{P}^{ij}$ defined by Eq. (29) may not necessarily reflect this error.

We have also encountered cases where a single station measurement MGEKF was divergent in the radial direction to the target, but no such difficulties have appeared when data from two geographically disparate stations was used. One way of generating such a divergent case was to decrease the maneuver colored noise autocorrelation parameter α to $\frac{1}{20}$ or below. We note that values of this parameter below $\frac{1}{20}$ correspond to slower maneuvers, a commonly encountered situation.

VI. Conclusions

This paper describes residual-based techniques for solving the radar track association problem for bearings-only measurements. The association between the tracks at two stations can be determined by examining the residuals of a bank of MGEKFs. Once this association is established, an algebraic parity test can find the correspondence between tracks at other stations and targets tracked by the first two stations.

One may ask why detection filters are necessary: Why not do everything with algebraic parity tests? Although the detection filtering step is not strictly necessary, it does improve the quality of the track associations because the state estimates constructed from two widely separated stations are so much more accurate than the estimates from a single station.

To ensure the quality of the estimates from the MGEKFs, one could delay the algebraic parity testing steps for associating tracks from additional stations. If these parity tests are replaced with additional detection filter banks until the estimates before and after including a new station's measurements are sufficiently close, then the fidelity of the estimates can be guaranteed.

Acknowledgments

This research was supported in part by the Air Force Office of Scientific Research under Grant F49620-00-1-0154 and by Sandia Laboratory under U.S. Department of Energy Grant LH-1376.

References

[1]Song, T. L., and Speyer, J. L., "A Stochastic Analysis of a Modified Gain Extended Kalman Filter with Applications to Estimation with Bearings Only Measurements," *IEEE Transactions on Automatic Control*, Vol. AC-30, No. 10, 1985, pp. 940–949.

[2]Singer, R. A., and Kanyuck, A. J., "Computer Control of Multiple Site Track Correlation," *Automatica*, Vol. 7, No. 4, 1971, pp. 455–463.

[3]Bar-Shalom, Y., "On the Track-to-Track Correlation Problem," *IEEE Transactions on Automatic Control*, Vol. AC-26, No. 2, 1981, pp. 571, 572.

[4]Bar-Shalom, Y., and Fortmann, T. E., *Tracking and Data Association*, Vol. 179, Mathematics in Science and Engineering, Academic Press, Boston, 1988, pp. 217–265, 266–272.

[5]Bar-Shalom, Y., and Campo, L., "The Effect of the Common Process Noise on the Two-Sensor Fused-Track Covariance," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-22, No. 6, 1986, pp. 803–805.

[6]Pao, L. Y., "Multisensor Multitarget Mixture Reduction Algorithms for Tracking," *Journal of Guidance, Control, and Dynamics*, Vol. 17, No. 6, 1994, pp. 1205–1211.

[7]Pao, L. Y., "Measurement Reconstruction Approach for Distributed Multisensor Fusion," *Journal of Guidance, Control, and Dynamics*, Vol. 19, No. 4, 1996, pp. 842–847.

[8]Fortmann, T. E., and Bar-Shalom, Y., "Sonar Tracking of Multiple Targets Using Joint Probabalistic Data Association," *Journal of Oceanic Engineering*, Vol. OE-8, No. 3, 1983, pp. 173–183.

[9]Reid, D. B., "An Algorithm for Tracking Multiple Targets," *IEEE Transactions on Automatic Control*, Vol. AC-24, No. 6, 1979, pp. 843–854.

Appendix K

"Exact Statistical Solution of Pseudorange Equations"

Jonathan D. Wolfe and Jason L. Speyer,

Proceedings of the ION GPS 2001 and to be published in *The Journal of the Institute of Navigation*

Exact Statistical Solution of Pseudorange Equations

Jonathan D. Wolfe *University of California, Los Angeles*
Jason L. Speyer *University of California, Los Angeles*

BIOGRAPHIES

Jonathan D. Wolfe is a Ph.D. candidate in the Mechanical and Aerospace Engineering Department at the University of California, Los Angeles.

Jason L. Speyer is a professor in the Mechanical and Aerospace Engineering Department at the University of California, Los Angeles. He received his Ph.D. degree in applied mathematics at Harvard University, Cambridge. Prof. Speyer is a Fellow of both the American Institute of Aeronautics and Astronautics and the Institute of Electrical and Electronics Engineers.

ABSTRACT

Although the exact GPS solution proposed by Bancroft is nonlinear, it may be manipulated into a linear form when 5 or more satellites are visible. This linear form is exact, as opposed to the linear solution obtained via repeated linearization in the iterated least squares (ILS) method. By virtue of this exactness, the solution of the linear form is always the true user position, while the ILS may converge to an incorrect solution (this is especially common when the GPS user is in space). When the measured pseudoranges are noisy, the linear structure ensures that the position estimate will converge to the correct value and that the error covariance of the estimate is known, guarantees that have not been found for nonlinear estimators that use the Bancroft solution directly. The conversion to the linear form excludes information present in a single scalar nonlinear measurement equation. We demonstrate several procedures for refining the linear estimate with this remaining information. In addition, we show that the methodology developed for direct GPS solutions can be applied to create linear direct methods for differential GPS problems.

1 INTRODUCTION

The purpose of the NAVSTAR global positioning system (GPS) is to allow a user to accurately determine their three-dimensional position. The system consists of a constellation of satellites, each of which broadcasts a predetermined time-varying code, known in advance by the user. The user can thus calculate the delay between the broadcast time from each satellite and the time of reception, which translates into a pseudorange from the receiver to that satellite. Since the positions of the GPS satellites are accurately known, these pseudoranges can be used to triangulate the user position. Because the user's clock does not align precisely with the clocks on the GPS satellites, the measured pseudoranges are not true ranges, and therefore the error between the user clock and GPS time must be estimated in order to accurately determine the user position.

The pseudoranges are nonlinear functions of four unknowns: the user position in three-space and the user clock error. Determining the unknowns therefore requires the pseudoranges from at least four non-coplanar GPS satellites. In view of this restriction, the orbits of the satellites in the GPS constellation were designed so that most earthbound users view at least four GPS satellites at any given time.

The most commonly used method for determining the unknowns is an iterated least squares (ILS) method, e.g. the one in Ref. [1]. This method amounts to a gradient search procedure to minimize the error between the measured pseudoranges and pseudoranges constructed from the estimated unknowns.

If the pseudoranges are noiseless, one may solve the pseudorange equations directly for the unknowns, as first reported by Bancroft [2]. This *direct method* was later refined and adapted to the case of noisy measurements [3, 4, 5, 6]. The estimates obtained by each of these methods all depend on the solution to a quadratic equation.

In this paper, we manipulate the measurement equations into a linear form, which is not dependent on

approximate knowledge of the user position. It is thus a direct method, but one that does not require the solution of a set of nonlinear measurement equations. Although Leva [7] has previously shown that such a linear structure existed, he used it solely for the purpose of determining when unique solutions to the GPS equations existed. In this work, we exploit the linear structure to obtain better estimates for systems with measurement noise.

The notation is as follows. The column vector $\mathbf{S}^{(i)}$ denotes the position of the ith satellite in an inertial reference frame, and the user position in the same reference frame is a column vector denoted by $\mathbf{x}$. The error between the user's clock and GPS time is denoted by Δt. This clock error causes a bias of $c\Delta t$ in all of the pseudoranges, where c denotes the speed of light. $\mathbf{X} \triangleq \begin{bmatrix}\mathbf{x}^T & c\Delta t\end{bmatrix}^T$ denotes the vector of unknowns for the GPS problem. The operator $< \cdot, \cdot >$ denotes the Euclidean inner product, and the operator $||\cdot||$ indicates the Euclidean norm. $E[\cdot]$ is the expectation operator. The number of GPS satellites visible to the user is m.

A review of the ILS algorithm appears in the next section, followed by a review of the direct solution in Section 3. Section 4 describes our linear direct method. An extension of this method is made to the differential GPS problem in Section 5. Because the linear direct method does not use all of the information present in the measurements, we propose several methods in Section 6 to improve upon the linear direct method's estimate by using the information present in the scalar nonlinear measurement equation that was excised in the construction of the linear direct method. Section 7 compares the accuracy of several direct methods operating on a simulated data set, and Section 8 concludes the paper.

2 REVIEW OF THE ILS ALGORITHM

For the ith visible GPS satellite, the measured range from the user to the satellite is given by

$$\tilde{\rho}^{(i)} = ||\mathbf{S}^{(i)} - \mathbf{x}|| + c\Delta t + \eta^{(i)}. \tag{1}$$

Let us denote $f(\mathbf{S}^{(i)}, \mathbf{X}) \triangleq ||\mathbf{S}^{(i)} - \mathbf{x}|| + c\Delta t$. When equation (1) is linearized about nominal values of $\tilde{\rho}_*^{(i)}$, $c\Delta t_*$ and $\mathbf{x}_*$, it becomes

$$\delta\tilde{\rho}^{(i)} = \mathbf{F}(\mathbf{S}^{(i)}, \mathbf{X})\delta\mathbf{X} + \eta^{(i)}, \tag{2}$$

where

$$\delta\tilde{\rho}^{(i)} = \tilde{\rho}^{(i)} - \tilde{\rho}_*^{(i)} = \tilde{\rho}^{(i)} - f(\mathbf{S}^{(i)}, \mathbf{X}_*), \tag{3}$$

$$\delta\mathbf{X} = \mathbf{X} - \mathbf{X}_*, \tag{4}$$

and

$$\mathbf{F}(\mathbf{S}^{(i)}, \mathbf{X}) \triangleq \frac{\partial f(\mathbf{S}^{(i)}, \mathbf{X})}{\partial \mathbf{X}}\Big|_{\mathbf{X}=\mathbf{X}_*}. \tag{5}$$

The ILS algorithm proceeds as follows:

Algorithm 2.1 (ILS).

1. *Set j=0. Let $\hat{\mathbf{X}}_0$ be the initial estimate of $\mathbf{X}$.*
2. *Set a convergence tolerance ϵ.*
3. *Let $\mathbf{W}$ be the covariance of the noise vector $\eta \triangleq \begin{bmatrix}\eta^{(1)} & \eta^{(2)} & \dots & \eta^{(m)}\end{bmatrix}^T$.*
4. *Compute $\mathbf{F}(\mathbf{S}^{(i)}, \hat{\mathbf{X}}_j), \quad i = 1, 2, \dots, m.$*
5. *Use the measurements $\rho^{(1)}, \rho^{(2)}, \dots, \rho^{(m)}$ and $\hat{\mathbf{X}}_j$ to compute $\delta\rho^{(1)}, \delta\rho^{(2)}, \dots, \delta\rho^{(m)}$ via equation (3).*
6. *Find the weighted least squares estimate of $\delta\mathbf{X}$ using the formula*

$$\delta\hat{\mathbf{X}} = (\mathbf{F}^T\mathbf{W}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{W}^{-1}\delta\rho, \tag{6}$$

where

$$\boldsymbol{\rho} \triangleq \begin{bmatrix}\rho^{(1)} \\ \rho^{(2)} \\ \vdots \\ \rho^{(m)}\end{bmatrix}, \qquad \mathbf{F} \triangleq \begin{bmatrix}\mathbf{F}(\mathbf{S}^{(1)}, \hat{\mathbf{X}}_j) \\ \mathbf{F}(\mathbf{S}^{(2)}, \hat{\mathbf{X}}_j) \\ \vdots \\ \mathbf{F}(\mathbf{S}^{(m)}, \hat{\mathbf{X}}_j)\end{bmatrix}.$$

7. *Update the estimate using the relation $\hat{\mathbf{X}}_{j+1} = \hat{\mathbf{X}}_j + \delta\hat{\mathbf{X}}$.*
8. *If $|\delta\mathbf{X}| < \epsilon$, the estimates have converged to within the tolerance ϵ and the algorithm can stop. If not, let $j = j + 1$ and return to step 4.*

Note that the initial guess $\hat{\mathbf{X}}_0$ and the convergence tolerance ϵ determine the number of iterations that the algorithm requires. If $\hat{\mathbf{X}}_0$ is a bad guess, the algorithm may converge very slowly.

3 REVIEW OF DIRECT METHODS FOR NOISY MEASUREMENTS

This section presents a brief review of the Improved Direct Solution (IDS) developed by Biton, *et al.* [6], which is an adjustment of Bancroft's direct method [2] to handle noisy data. Begin with the equation for the measured pseudorange from the user to the ith satellite

$$\tilde{\rho}^{(i)} = ||\mathbf{S}^{(i)} - \mathbf{x}|| + c\Delta t + \eta^{(i)}, \tag{7}$$

or equivalently

$$||\mathbf{S}^{(i)} - \mathbf{x}|| = \tilde{\rho}^{(i)} - c\Delta t - \eta^{(i)}. \tag{8}$$

Squaring this equation and rearranging terms yields

$$-2 < \mathbf{S}^{(i)}, \mathbf{x} > + \chi + 2\tilde{\rho}^{(i)}c\Delta t - (\eta^{(i)})^2 + 2\tilde{\rho}^{(i)}\eta^{(i)} - 2\eta^{(i)}c\Delta t = (\tilde{\rho}^{(i)})^2 - ||\mathbf{S}^{(i)}||^2, \tag{9}$$

where

$$\chi \triangleq ||\mathbf{x}||^2 - (c\Delta t)^2. \quad (10)$$

Typical values of the quantities in equation (9) are

$$||\mathbf{S}^{(i)}|| \approx 10^7 m, \quad ||x|| \approx 10^5 m, \quad c\Delta t \approx 10^5 m,$$
$$\eta^{(i)} \approx 1m, \quad \tilde{\rho}^{(i)} \approx 10^7 m.$$

If we ignore all terms in (9) that are smaller than 10^8 meters, we have

$$-2 < \mathbf{S}^{(i)}, \mathbf{x} > + \chi + 2\tilde{\rho}^{(i)} c\Delta t + 2\tilde{\rho}^{(i)}\eta^{(i)} = (\tilde{\rho}^{(i)})^2 - ||\mathbf{S}^{(i)}||^2. \quad (11)$$

We can assemble a vector equation by applying (11) to the measurements from every satellite:

$$\mathbf{HX} + \mathbf{G}\boldsymbol{\eta} = \mathbf{R}_a + \chi \mathbf{R}_b, \quad (12)$$

where

$$\mathbf{H} \triangleq 2\begin{bmatrix} -(\mathbf{S}^{(1)})^T & \tilde{\rho}^{(1)} \\ -(\mathbf{S}^{(2)})^T & \tilde{\rho}^{(2)} \\ \vdots & \vdots \\ -(\mathbf{S}^{(m)})^T & \tilde{\rho}^{(m)} \end{bmatrix}, \quad (13)$$

$$\mathbf{R}_a \triangleq \begin{bmatrix} (\tilde{\rho}^{(1)})^2 - ||\mathbf{S}^{(1)}||^2 \\ (\tilde{\rho}^{(2)})^2 - ||\mathbf{S}^{(2)}||^2 \\ \vdots \\ (\tilde{\rho}^{(m)})^2 - ||\mathbf{S}^{(m)}||^2 \end{bmatrix}, \quad \mathbf{R}_b \triangleq \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}, \quad (14)$$

$$\mathbf{G} \triangleq \mathrm{diag}\{2\tilde{\rho}^{(1)}, 2\tilde{\rho}^{(2)}, \ldots, 2\tilde{\rho}^{(m)}\}, \quad (15)$$

$$\boldsymbol{\eta} \triangleq \begin{bmatrix} \eta^{(1)} & \eta^{(2)} & \ldots & \eta^{(m)} \end{bmatrix}^T. \quad (16)$$

Since $\mathbf{G}$ is invertible, equation (12) can be rewritten as

$$\mathbf{G}^{-1}\mathbf{HX} + \boldsymbol{\eta} = \mathbf{G}^{-1}\mathbf{R}_a + \chi \mathbf{G}^{-1}\mathbf{R}_b. \quad (17)$$

Suppose that $\boldsymbol{\eta}$ were a zero mean Gaussian vector with covariance $\mathbf{V}$. If χ was a known quantity, and if the *a priori* estimate of $\mathbf{X}$ was the zero vector, and if the error covariance of this initial estimate was infinite, then

$$\hat{\mathbf{X}} = [(\mathbf{G}^{-1}\mathbf{H})^T \mathbf{V}^{-1} (\mathbf{G}^{-1}\mathbf{H})]^{-1} \cdot (\mathbf{G}^{-1}\mathbf{H})^T \mathbf{G}^{-1} (\mathbf{R}_a + \chi \mathbf{R}_b) \quad (18)$$

is the least squares estimate of $\mathbf{X}$. Substitution of the elements of $\hat{\mathbf{X}}$ in (18) into the definition of χ (10) results in a quadratic equation in the unknown χ (the coefficients of this quadratic equation are independent of $\mathbf{X}$ and χ, hence the nomenclature "direct solution"). Substituting the two solutions of this equation into (18) gives two candidates for $\hat{\mathbf{X}}$. Only one of these candidates will satisfy the measurement equation (12).

Note that for the noiseless case, the estimation equation (18) simplifies to

$$\hat{\mathbf{X}} = [\mathbf{H}^T\mathbf{H}]^{-1}\mathbf{H}^T(\mathbf{R}_a + \chi \mathbf{R}_b), \quad (19)$$

which is the estimation equation used in the original Bancroft method.

4 THE LINEAR DIRECT EQUATIONS

The only nonlinear term in the vector measurement equation (17) is χ. We demonstrate the means to remove this nonlinearity below.

The matrix $\mathbf{G}^{-1}\mathbf{R}_b$ is rank 1. Hence, a rank $(m-1)$ matrix $\mathbf{E}$ exists such that $\mathbf{E}\mathbf{G}^{-1}\mathbf{R}_b = \mathbf{0}$. In fact, there are an infinite number of such annihilator matrices, with

$$\mathbf{E} = \begin{bmatrix} 1 & -1 & 0 & & 0 \\ 1 & 0 & -1 & & 0 \\ & & & \ddots & \\ 1 & 0 & 0 & & -1 \end{bmatrix} \mathbf{G}$$

being an obvious example. Multiplying the GPS measurement equation (17) on the left by the annihilator $\mathbf{E}$ creates $(m-1)$ completely linear exact GPS measurement equations:

$$\mathbf{E}\mathbf{G}^{-1}\mathbf{HX} + \mathbf{E}\boldsymbol{\eta} = \mathbf{E}\mathbf{G}^{-1}\mathbf{R}_a. \quad (20)$$

Remark 4.1. *If the noise vector $\boldsymbol{\eta}$ is assumed to be a zero mean Gaussian with covariance $\mathbf{V} \triangleq E[\boldsymbol{\eta}\boldsymbol{\eta}^T]$, the single epoch least squares solution of the linear measurement equation* (20) *that minimizes the cost function*

$$J = [\mathbf{E}\mathbf{G}^{-1}\mathbf{R}_a - \mathbf{E}\mathbf{G}^{-1}\mathbf{HX}]^T (\mathbf{E}\mathbf{V}\mathbf{E}^T)^{-1} \cdot [\mathbf{E}\mathbf{G}^{-1}\mathbf{R}_a - \mathbf{E}\mathbf{G}^{-1}\mathbf{HX}] + \mathbf{X}^T\mathbf{M}_0^{-1}\mathbf{X} \quad (21)$$

is given by

$$\hat{\mathbf{X}}_0 = [\mathbf{H}^T(\mathbf{G}^{-1})^T\mathbf{E}^T(\mathbf{E}\mathbf{V}\mathbf{E}^T)^{-1}\mathbf{E}\mathbf{G}^{-1}\mathbf{H} + \mathbf{M}_0^{-1}]^{-1} \cdot \mathbf{H}^T(\mathbf{G}^{-1})^T\mathbf{E}^T(\mathbf{E}\mathbf{V}\mathbf{E}^T)^{-1} \cdot \left(\mathbf{E}\mathbf{G}^{-1}\mathbf{R}_a - \mathbf{E}\mathbf{G}^{-1}\mathbf{H}\bar{\mathbf{X}}_0\right), \quad (22)$$

where $\bar{\mathbf{X}}_0$ is the a priori *state estimate and $\mathbf{M}_0$ is the (potentially infinite)* a priori *error covariance. The error covariance $\mathbf{P}_0 \triangleq E[(\mathbf{X} - \hat{\mathbf{X}}_0)(\mathbf{X} - \hat{\mathbf{X}}_0)^T]$ of the estimate is*

$$\mathbf{P}_0 = \left(\mathbf{H}^T(\mathbf{G}^{-1})^T\mathbf{E}^T(\mathbf{E}\mathbf{V}\mathbf{E}^T)^{-1} \cdot \mathbf{E}\mathbf{G}^{-1}\mathbf{H} + \mathbf{M}_0^{-1}\right)^{-1}. \quad (23)$$

Note that this error covariance will typically be larger than that for an ILS estimate, since the projected measurements have effectively double the noise of the original measurements. Hence, using the estimate obtained in (22) *as the initial guess $\hat{\mathbf{X}}_0$ in the ILS algorithm 2.1 would be a good strategy. Since $\hat{\mathbf{X}}_0$ is very close to the true position, the ILS algorithm should converge in a step or two. Section 6 derives a similar method that gives even better results.*

If a state model is available, a Kalman filter can be constructed for estimating the vector $\mathbf{X}$. Assuming that the measurement noise sequence $\boldsymbol{\eta}$ is an independent

and identically distributed (i.i.d.) zero mean Gaussian random sequence with covariance $\mathbf{V}$*, equations* (22) *and* (23) *describe the update equations for such a Kalman filter. Note that an extended Kalman filter is not required, as all of the measurement equations are linear.*

5 EXACT LINEAR SOLUTION OF NOISY DIFFERENTIAL GPS EQUATIONS

Suppose that we are interested in knowing the distance $\Delta\mathbf{x} \triangleq \mathbf{x}_2 - \mathbf{x}_1$ between two receivers located at the positions $\mathbf{x}_1$ and $\mathbf{x}_2$. Let $c\Delta t_1$ be the clock bias of first receiver, and let $c\Delta t_2$ be the clock bias of the second. Define $c\Delta t_{12}$ as the difference between the two clock biases, $c\Delta t_{12} \triangleq c\Delta t_2 - c\Delta t_1$.

For each satellite i that is visible to both receivers, there are two measurements available: $\tilde{\rho}_1^{(i)} \triangleq ||\mathbf{S}^{(i)} - \mathbf{x}_1|| + c\Delta t_1 + \eta_1^{(i)}$ and $\delta\tilde{\rho}^{(i)} \triangleq ||\mathbf{S}^{(i)} - \mathbf{x}_2|| - ||\mathbf{S}^{(i)} - \mathbf{x}_1|| + c\Delta t_{12} + \eta_{12}^{(i)}$. Note that $\eta_1^{(i)}$ and $\eta_{12}^{(i)}$ are correlated, but the magnitude of $\eta_{12}^{(i)}$ is less than that of $\eta_1^{(i)}$ due to the elimination of common mode errors from the differential measurement. If $\delta\tilde{\rho}^{(i)}$ is a differential carrier phase measurement, then $\eta_{12}^{(i)}$ should be very small compared to $\eta_1^{(i)}$, since carrier phase multipath is much smaller than code multipath.

Proposition 5.1. *The distance between the receivers* $\Delta\mathbf{x}$ *satisfies the following equation for each visible satellite* i*:*

$$\begin{aligned}
&-2 < \mathbf{S}^{(i)}, \Delta\mathbf{x} > + ||\mathbf{x}_2||^2 - ||\mathbf{x}_1||^2 + 2\tilde{\rho}_1^{(i)} c\Delta t_{12} + \\
&2\delta\tilde{\rho}^{(i)} c\Delta t_1 + 2\delta\tilde{\rho}^{(i)} c\Delta t_{12} - 2c\Delta t_1 c\Delta t_{12} - (c\Delta t_{12})^2 + \\
&2\tilde{\rho}^{(i)}\eta_{12}^{(i)} - 2c\Delta t_{12}\eta_1^{(i)} - 2\eta_1^{(i)}\eta_{12}^{(i)} - 2c\Delta t_1\eta_{12}^{(i)} + 2\delta\tilde{\rho}^{(i)}\eta_{12}^{(i)} - \\
&(\eta_{12}^{(i)})^2 - 2c\Delta t_{12}\eta_{12}^{(i)} = 2\tilde{\rho}_1^{(i)}\delta\tilde{\rho}^{(i)} + (\delta\tilde{\rho}^{(i)})^2. \quad (24)
\end{aligned}$$

Proof. Begin by noting the identity

$$\begin{aligned}
&||\mathbf{S}^{(i)} - \mathbf{x}_2||^2 = \\
&(||\mathbf{S}^{(i)} - \mathbf{x}_1|| + (||\mathbf{S}^{(i)} - \mathbf{x}_2|| - ||\mathbf{S}^{(i)} - \mathbf{x}_1||))^2. \quad (25)
\end{aligned}$$

Expanding the quadratic terms on both sides yields

$$\begin{aligned}
&||\mathbf{S}^{(i)}||^2 - 2 < \mathbf{S}^{(i)}, \mathbf{x}_2 > + ||\mathbf{x}_2||^2 = \\
&\quad ||\mathbf{S}^{(i)}||^2 - 2 < \mathbf{S}^{(i)}, \mathbf{x}_1 > + ||\mathbf{x}_1||^2 + \\
&\quad 2||\mathbf{S}^{(i)} - \mathbf{x}_1||(||\mathbf{S}^{(i)} - \mathbf{x}_2|| - ||\mathbf{S}^{(i)} - \mathbf{x}_1||) + \\
&\quad (||\mathbf{S}^{(i)} - \mathbf{x}_2|| - ||\mathbf{S}^{(i)} - \mathbf{x}_1||)^2. \quad (26)
\end{aligned}$$

Now substitute the definitions of $\Delta\mathbf{x}$, $\tilde{\rho}_1^{(i)}$ and $\delta\tilde{\rho}^{(i)}$ into the above expression and rearrange:

$$\begin{aligned}
&-2 < \mathbf{S}^{(i)}, \Delta\mathbf{x} > + ||\mathbf{x}_2||^2 - ||\mathbf{x}_1||^2 = \\
&\quad 2(\tilde{\rho}_1^{(i)} - c\Delta t_1 - \eta_1^{(i)})(\delta\tilde{\rho}^{(i)} - c\Delta t_{12} - \eta_{12}^{(i)}) + \\
&\quad (\delta\tilde{\rho}^{(i)} - c\Delta t_{12} - \eta_{12}^{(i)})^2. \quad (27)
\end{aligned}$$

Let us expand the terms on the right hand side containing the clock biases, giving us

$$\begin{aligned}
&-2 < \mathbf{S}^{(i)}, \Delta\mathbf{x} > + ||\mathbf{x}_2||^2 - ||\mathbf{x}_1||^2 = \\
&\quad 2(\tilde{\rho}_1^{(i)} - \eta_1^{(i)})(\delta\tilde{\rho}^{(i)} - c\Delta t_{12} - \eta_{12}^{(i)}) - \\
&\quad 2c\Delta t_1\delta\tilde{\rho}^{(i)} + 2c\Delta t_1 c\Delta t_{12} + 2c\Delta t_1\eta_{12}^{(i)} + \\
&(\delta\tilde{\rho}^{(i)} - \eta_{12}^{(i)})^2 - 2c\Delta t_{12}(\delta\tilde{\rho}^{(i)} - \eta_{12}^{(i)}) + (c\Delta t_{12})^2, \quad (28)
\end{aligned}$$

which can be rearranged as (24). □

We can assemble a vector equation by applying (24) to the measurements from every satellite:

$$\mathbf{H}_d\Delta\mathbf{X} + \mathbf{G}_d\eta_d + \mathbf{R}_{cd} = \mathbf{R}_{ad} + \chi_d\mathbf{R}_{bd}, \quad (29)$$

where

$$\mathbf{H}_d \triangleq 2\begin{bmatrix} -(\mathbf{S}^{(1)})^T & \delta\tilde{\rho}^{(1)} & (\tilde{\rho}^{(1)} + \delta\tilde{\rho}^{(1)}) \\ -(\mathbf{S}^{(2)})^T & \delta\tilde{\rho}^{(2)} & (\tilde{\rho}^{(2)} + \delta\tilde{\rho}^{(2)}) \\ \vdots & \vdots & \vdots \\ -(\mathbf{S}^{(m)})^T & \delta\tilde{\rho}^{(m)} & (\tilde{\rho}^{(m)} + \delta\tilde{\rho}^{(m)}) \end{bmatrix}, \quad (30)$$

$$\Delta\mathbf{X} \triangleq \begin{bmatrix} \Delta\mathbf{x} \\ c\Delta t_1 \\ c\Delta t_{12} \end{bmatrix}, \qquad \eta_d \triangleq \begin{bmatrix} \eta_{12}^{(1)} \\ \eta_{12}^{(2)} \\ \vdots \\ \eta_{12}^{(m)} \end{bmatrix} \quad (31)$$

$$\begin{aligned}
\mathbf{G}_d \triangleq 2\,\mathrm{diag}\{&(\tilde{\rho}^{(1)} + \delta\tilde{\rho}^{(1)}), (\tilde{\rho}^{(2)} + \delta\tilde{\rho}^{(2)}), \ldots \\
&, (\tilde{\rho}^{(m)} + \delta\tilde{\rho}^{(m)})\}, \quad (32)
\end{aligned}$$

$$\begin{aligned}
R_{cd}^{(i)} \triangleq &-2c\Delta t_{12}\eta_1^{(i)} - 2\eta_1^{(i)}\eta_{12}^{(i)} - 2c\Delta t_1\eta_{12}^{(i)} - \\
&(\eta_{12}^{(i)})^2 - 2c\Delta t_{12}\eta_{12}^{(i)}, \qquad i = 1, 2, \ldots, m \quad (33)
\end{aligned}$$

$$\mathbf{R}_{cd} \triangleq \begin{bmatrix} R_{cd}^{(1)} \\ R_{cd}^{(2)} \\ \vdots \\ R_{cd}^{(m)} \end{bmatrix}, \qquad \mathbf{R}_{bd} \triangleq \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}, \quad (34)$$

$$\mathbf{R}_{ad} \triangleq \begin{bmatrix} 2\tilde{\rho}_1^{(1)}\delta\tilde{\rho}^{(1)} + (\delta\tilde{\rho}^{(1)})^2 \\ 2\tilde{\rho}_1^{(2)}\delta\tilde{\rho}^{(2)} + (\delta\tilde{\rho}^{(2)})^2 \\ \vdots \\ 2\tilde{\rho}_1^{(m)}\delta\tilde{\rho}^{(m)} + (\delta\tilde{\rho}^{(m)})^2 \end{bmatrix}, \quad (35)$$

$$\chi_d \triangleq ||\mathbf{x}_2||^2 - ||\mathbf{x}_1||^2 - 2c\Delta t_1 c\Delta t_{12} - (c\Delta t_{12})^2. \quad (36)$$

Proposition 5.2. $\Delta\mathbf{X}$*, the solution to the vector equation* (29)*, satisfies the rank* $(m-1)$ *vector equation*

$$\begin{aligned}
&\mathbf{E}_d\mathbf{G}_d^{-1}\mathbf{H}_d\Delta\mathbf{X} + \mathbf{E}_d\eta_d + \mathbf{E}_d\mathbf{G}_d^{-1}\mathbf{R}_{cd} = \\
&\qquad \mathbf{E}_d\mathbf{G}_d^{-1}\mathbf{R}_{ad}, \quad (37)
\end{aligned}$$

which is linear in the unknowns $\Delta\mathbf{x}$*,* $c\Delta t_1$*, and* $c\Delta t_{12}$*.*

Proof. As for the standard GPS case, there exists a rank $(m-1)$ left annihilator $\mathbf{E}_d$ to the rank 1 vector $\mathbf{G}_d^{-1}\mathbf{R}_{bd}$. Multiply (29) on the left by $\mathbf{E}_d\mathbf{G}_d^{-1}$ to obtain (37). □

Remark 5.3 (Approximate Linearity). *Equation* (37) *is already linear in the unknowns* $\Delta\mathbf{x}$, $c\Delta t_{12}$, *and* $c\Delta t_1$. *If we make some mild approximations, we can remove the nonlinearities in the noise terms. The elements in* (37) *typically are of the following sizes:*

$$||\mathbf{S}^{(i)}|| \approx 10^7 m, \qquad \tilde{\rho}_1^{(i)} \approx 10^7 m, \qquad c\Delta t_1 \approx 10^5 m,$$
$$c\Delta t_{12} \approx 10^5 m, \qquad \eta_1^{(i)} \approx 1m, \qquad \eta_{12}^{(i)} \approx 1m,$$

$$0m < ||\Delta x|| < 10^7 m, \qquad 0m < ||\delta\tilde{\rho}^{(i)}|| < 10^7 m.$$

If we choose an unitary annihilator (i.e. $\mathbf{E}_d\mathbf{G}_d^{-1}\mathbf{R}_{bd} = \mathbf{0}$, $\mathbf{E}_d\mathbf{E}_d^T = \mathbf{I}_{m-1}$. *Such an annihilator always exists.) and ignore all terms in* (37) *that are smaller than* 10 *meters, we have*

$$\mathbf{E}_d\mathbf{G}_d^{-1}\mathbf{H}_d\Delta\mathbf{X} + \mathbf{E}_d\boldsymbol{\eta}_d = \mathbf{E}_d\mathbf{G}_d^{-1}\mathbf{R}_{ad}, \tag{38}$$

which is linear in the unknown noises as well as in $\Delta\mathbf{x}$, $c\Delta t_{12}$, *and* $c\Delta t_1$. *Note that this formulation holds even for very long baselines. In contrast, the linearization conventionally used for the measurement equation introduces significant errors when the distance between the two antennas is larger than 100 kilometers [8]. It should be noted that terrestrial differential GPS over long baselines is still subject to larger noises than those associated with shorter baselines, due to noncommon mode ionospheric errors.*

6 NONLINEAR CORRECTION

In the course of the derivation of the linear exact solution, some of the information in the measurement vector $\mathbf{G}^{-1}\mathbf{R}_a$ is not used, namely the part of $\mathbf{G}^{-1}\mathbf{R}_a$ that is orthogonal to the projector $\mathbf{E}$. This information can be recovered by applying $\mathbf{R}_b^T(\mathbf{G}^{-1})^T$, the orthogonal complement of $\mathbf{E}$, to the nonlinear measurement equation (17), which yields the following scalar nonlinear equation:

$$\breve{\mathbf{H}}\mathbf{X} + \breve{\mathbf{G}}\boldsymbol{\eta} = \breve{\mathbf{R}}_a + \breve{\mathbf{R}}_b\chi, \tag{39}$$

where

$$\breve{\mathbf{R}}_a \triangleq \mathbf{R}_b^T(\mathbf{G}^{-1})^T\mathbf{G}^{-1}\mathbf{R}_a, \tag{40}$$
$$\breve{\mathbf{R}}_b \triangleq \mathbf{R}_b^T(\mathbf{G}^{-1})^T\mathbf{G}^{-1}\mathbf{R}_b, \tag{41}$$
$$\breve{\mathbf{H}} \triangleq \mathbf{R}_b^T(\mathbf{G}^{-1})^T\mathbf{G}^{-1}\mathbf{H}, \qquad \breve{\mathbf{G}} \triangleq \mathbf{R}_b^T(\mathbf{G}^{-1})^T. \tag{42}$$

Remark 6.1 (Independence of measurement noise). *Note that if the measurement noise* $\boldsymbol{\eta}$ *is assumed to be composed of independent and identically distributed zero mean Gaussians* (i.e. $E[\boldsymbol{\eta}] = \mathbf{0}$, $E[\boldsymbol{\eta}\boldsymbol{\eta}] = \sigma^2\mathbf{I}$) *then* $\mathbf{R}_b^T(\mathbf{G}^{-1})^T\boldsymbol{\eta}$ *is uncorrelated to the noise in the measurement equation* (20)*:*

$$E[\mathbf{E}\boldsymbol{\eta}(\mathbf{R}_b^T(\mathbf{G}^{-1})^T\boldsymbol{\eta})^T] = E[\mathbf{E}\boldsymbol{\eta}\boldsymbol{\eta}^T\mathbf{G}^{-1}\mathbf{R}_b] \tag{43}$$
$$= \mathbf{E}(\sigma^2\mathbf{I})\mathbf{G}^{-1}\mathbf{R}_b \tag{44}$$
$$= \sigma^2\mathbf{E}\mathbf{G}^{-1}\mathbf{R}_b^T = \mathbf{0}. \tag{45}$$

This means in particular that $\mathbf{R}_b^T(\mathbf{G}^{-1})^T\boldsymbol{\eta}$ *is independent of* $\hat{\mathbf{X}}_0$, *the estimate based on the measurement equation* (20).

The independence of $\hat{\mathbf{X}}_0$ and the noise in (39) suggests several methods for adding a correction to the linear estimate based on the nonlinear part of the measurement equation. Some possible corrections are a linearization of the nonlinear equations, a minimum variance linear estimate based on the nonlinear part, a maximum likelihood estimate based on the nonlinear part, and a conditional mean estimate based on the nonlinear part. The following subsections detail each of these methods.

6.1 LINEARIZATION OF NONLINEAR PART

When linearized about $\hat{\mathbf{X}}_0$, equation (39) becomes

$$\tilde{\mathbf{H}}\mathbf{X} + \tilde{\boldsymbol{\eta}} = \breve{\mathbf{R}}_a, \tag{46}$$

where

$$\tilde{\mathbf{H}} \triangleq \{\breve{\mathbf{H}} - \breve{\mathbf{R}}_b\begin{bmatrix}\hat{\mathbf{x}}_0^T & -c\widehat{\Delta t}_0\end{bmatrix}\}, \tag{47}$$
$$\tilde{\boldsymbol{\eta}} \triangleq \mathbf{R}_b^T(\mathbf{G}^{-1})^T\boldsymbol{\eta}. \tag{48}$$

Since $\tilde{\boldsymbol{\eta}}$ is independent of $\hat{\mathbf{X}}_0$ the updated estimate that makes use of the linearized measurement equation (46) is calculated using the standard least squares update

$$\hat{\mathbf{X}} = \hat{\mathbf{X}}_0 + \mathbf{P}_0\tilde{\mathbf{H}}^T\big(\tilde{\mathbf{H}}\mathbf{P}_0\tilde{\mathbf{H}}^T + (\sigma^2\breve{\mathbf{R}}_b)^{-1}\big)\cdot \big(\breve{\mathbf{R}}_a - \mathbf{R}_b^T(\mathbf{G}^{-1})^T\mathbf{G}^{-1}\hat{\mathbf{R}}_a\big), \tag{49}$$

where $\hat{\mathbf{R}}_a$ is the measurement corresponding to $\hat{\mathbf{X}}_0$:

$$\hat{\mathbf{R}}_a \triangleq \begin{bmatrix} \big(||\mathbf{S}^1 - \hat{\mathbf{x}}_0||^2 + c\widehat{\Delta t}_0\big)^2 - ||\mathbf{S}^1||^2 \\ \big(||\mathbf{S}^2 - \hat{\mathbf{x}}_0||^2 + c\widehat{\Delta t}_0\big)^2 - ||\mathbf{S}^2||^2 \\ \vdots \\ \big(||\mathbf{S}^m - \hat{\mathbf{x}}_0||^2 + c\widehat{\Delta t}_0\big)^2 - ||\mathbf{S}^m||^2 \end{bmatrix}. \tag{50}$$

6.2 MAXIMUM LIKELIHOOD ESTIMATE USING NONLINEAR PART

The *a priori* probability density function (pdf) of $\mathbf{X}$ given by the linear estimate equations (22) and (23) is

$$p(\mathbf{X}) = \frac{1}{(2\pi)^2|\mathbf{P}_0|^{1/2}}\cdot \exp\{-\frac{1}{2}(\mathbf{X} - \hat{\mathbf{X}}_0)^T\mathbf{P}_0^{-1}(\mathbf{X} - \hat{\mathbf{X}}_0)\}. \tag{51}$$

Our objective is to maximize the joint probability density function of $\mathbf{X}$ and the measurement $\breve{\mathbf{R}}_a$. By Bayes' rule,

this joint pdf can be expressed in terms of the conditional pdf

$$p(\mathbf{X}, \breve{\mathbf{R}}_a) = p(\breve{\mathbf{R}}_a|\mathbf{X})p(\mathbf{X}).$$

The conditional pdf we require is determined by using the nonlinear measurement equation (39)

$$p(\breve{\mathbf{R}}_a|\mathbf{X}) = \frac{1}{\sqrt{2\pi}\breve{\sigma}} \cdot \exp\{-\frac{1}{2\breve{\sigma}^2}(\breve{\mathbf{R}}_a - \breve{\mathbf{H}}\mathbf{X} - \breve{\mathbf{R}}_b\mathbf{X}^T\mathbf{Q}\mathbf{X})^2\}, \quad (52)$$

where

$$\breve{\sigma}^2 \triangleq E[\tilde{\eta}\tilde{\eta}] = (\mathbf{G}^{-1}\mathbf{R}_b)^T E[\eta\eta^T](\mathbf{G}^{-1}\mathbf{R}_b) = (\mathbf{G}^{-1}\mathbf{R}_b)^T(\sigma^2\mathbf{I})(\mathbf{G}^{-1}\mathbf{R}_b) \quad (53)$$

and

$$\mathbf{Q} \triangleq \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}. \quad (54)$$

Then

$$p(\mathbf{X}, \breve{\mathbf{R}}_a) = \frac{1}{(2\pi)^{5/2}\breve{\sigma}|\mathbf{P}_0|^{1/2}} \cdot \exp\{-\frac{1}{2\breve{\sigma}^2}(\breve{\mathbf{R}}_a - \breve{\mathbf{H}}\mathbf{X} - \breve{\mathbf{R}}_b\mathbf{X}^T\mathbf{Q}\mathbf{X})^2 - \frac{1}{2}(\mathbf{X} - \hat{\mathbf{X}}_0)^T\mathbf{P}_0^{-1}(\mathbf{X} - \hat{\mathbf{X}}_0)\} \quad (55)$$

is the joint pdf that we must maximize with respect to $\mathbf{X}$. Clearly, maximizing $p(\mathbf{X}, \breve{\mathbf{R}}_a)$ above is equivalent to minimizing the function

$$f(\mathbf{X}) \triangleq \frac{1}{2\breve{\sigma}^2}(\breve{\mathbf{R}}_a - \breve{\mathbf{H}}\mathbf{X} - \breve{\mathbf{R}}_b\mathbf{X}^T\mathbf{Q}\mathbf{X})^2 + \frac{1}{2}(\mathbf{X} - \hat{\mathbf{X}}_0)^T\mathbf{P}_0^{-1}(\mathbf{X} - \hat{\mathbf{X}}_0). \quad (56)$$

The maximum likelihood estimate can thus be obtained by minimizing the function $f(\mathbf{X})$ expressed in equation (56). Note that f is not necessarily convex in $\mathbf{X}$, so this minimization problem may not be an easy one. Since f is a smooth function, it is convex if and only if its' Hessian is positive definite everywhere. The relevant quantities to determine convexity and solve for the optimum via a Newton-Raphson algorithm are

$$\frac{\partial}{\partial \mathbf{X}} f(\mathbf{X}) = -\frac{1}{\breve{\sigma}^2}(\breve{\mathbf{R}}_a - \breve{\mathbf{H}}\mathbf{X} - \breve{\mathbf{R}}_b\mathbf{X}^T\mathbf{Q}\mathbf{X}) \cdot (\breve{\mathbf{H}} + 2\breve{\mathbf{R}}_b\mathbf{X}^T\mathbf{Q}) + (\mathbf{X} - \hat{\mathbf{X}}_0)^T\mathbf{P}_0^{-1} \quad (57)$$

$$\frac{\partial^2}{\partial \mathbf{X}^2} f(\mathbf{X}) = -\frac{2}{\breve{\sigma}^2}\breve{\mathbf{R}}_b(\breve{\mathbf{R}}_a - \breve{\mathbf{H}}\mathbf{X} - \breve{\mathbf{R}}_b\mathbf{X}^T\mathbf{Q}\mathbf{X})\mathbf{Q} + \frac{1}{\breve{\sigma}^2}(\breve{\mathbf{H}} + 2\breve{\mathbf{R}}_b\mathbf{X}^T\mathbf{Q})^T(\breve{\mathbf{H}} + 2\breve{\mathbf{R}}_b\mathbf{X}^T\mathbf{Q}) \mathbf{P}_0^{-1}. \quad (58)$$

6.3 CONDITIONAL MEAN ESTIMATE USING NONLINEAR PART

In the previous subsection, the joint probability density function $p(\mathbf{X}, \breve{\mathbf{R}}_a)$ was calculated. The density function $p(\breve{\mathbf{R}}_a)$ is given by

$$p(\breve{\mathbf{R}}_a) = \int_{-\infty}^{\infty} p(\mathbf{X}, \breve{\mathbf{R}}_a)\, d\mathbf{X}, \quad (59)$$

and by Bayes' rule,

$$p(\mathbf{X}|\breve{\mathbf{R}}_a) = \frac{p(\mathbf{X}, \breve{\mathbf{R}}_a)}{p(\breve{\mathbf{R}}_a)}. \quad (60)$$

Then the conditional mean is

$$E[\mathbf{X}|\breve{\mathbf{R}}_a] = \int_{-\infty}^{\infty} \mathbf{X}\, p(\mathbf{X}|\breve{\mathbf{R}}_a)\, d\mathbf{X}. \quad (61)$$

This conditional mean estimate is the one we really want, although the above integrals are difficult to compute. Until these integrals are solved, the maximum likelihood solution from the last section remains the best viable alternative.

6.4 MINIMUM VARIANCE ESTIMATE USING NONLINEAR PART

Let us describe the random variable $\mathbf{X}$ as

$$\mathbf{X} = \hat{\mathbf{X}}_0 + \Delta\mathbf{X},$$

where $\hat{\mathbf{X}}_0$ is the *a priori* estimate generated using the linear part of the measurement equations and $\Delta\mathbf{X}$ is a zero-mean Gaussian random variable with covariance $\mathbf{P}_0$. Then the nonlinear measurement equation (39) can be expressed as

$$\breve{\mathbf{H}}(\hat{\mathbf{X}}_0 + \Delta\mathbf{X}) + \breve{\mathbf{R}}_b(\hat{\mathbf{X}}_0 + \Delta\mathbf{X})^T\mathbf{Q}(\hat{\mathbf{X}}_0 + \Delta\mathbf{X}) + \breve{\mathbf{G}}\eta = \breve{\mathbf{R}}_a, \quad (62)$$

or upon rearrangement as

$$\breve{\mathbf{H}}_1\Delta\mathbf{X} + \breve{\mathbf{R}}_b(\Delta\mathbf{X})^T\mathbf{Q}(\Delta\mathbf{X}) + \breve{\mathbf{G}}\eta = \mathbf{R}_1, \quad (63)$$

where

$$\breve{\mathbf{H}}_1 = \breve{\mathbf{H}} + 2\breve{\mathbf{R}}_b\hat{\mathbf{X}}_0^T\mathbf{Q}, \tag{64}$$

$$\mathbf{R}_1 = \breve{\mathbf{R}}_a - \breve{\mathbf{H}}\hat{\mathbf{X}}_0 - \breve{\mathbf{R}}_b\hat{\mathbf{X}}_0^T\mathbf{Q}\hat{\mathbf{X}}_0. \tag{65}$$

Given this measurement equation and the *a priori* distribution of $\Delta\mathbf{X}$, the linear minimum variance estimate of $\Delta\mathbf{X}$ is

$$\widehat{\Delta\mathbf{X}} = E[(\Delta\mathbf{X})\mathbf{R}_1^T]E[\mathbf{R}_1\mathbf{R}_1^T]^{-1}\mathbf{R}_1, \tag{66}$$

as demonstrated in [9]. Since we have already shown that $\Delta\mathbf{X}$ is independent of the measurement noise $\breve{\mathbf{G}}\eta$, the expectations in the above equation are readily calculated:

$$E[(\Delta\mathbf{X})\mathbf{R}_1^T] = \mathbf{P}_0\breve{\mathbf{H}}^T \tag{67}$$

$$\begin{aligned} E[\mathbf{R}_1\mathbf{R}_1^T] = \breve{\mathbf{R}}_b^2\{&3(\mathbf{P}_{11}^2 + \mathbf{P}_{22}^2 + \mathbf{P}_{33}^2 + \mathbf{P}_{44}^2) + \\ &4\mathbf{P}_{12}^2 + 2\mathbf{P}_{11}\mathbf{P}_{22} + 4\mathbf{P}_{13}^2 + \\ &2\mathbf{P}_{11}\mathbf{P}_{33} + 4\mathbf{P}_{23}^2 + 2\mathbf{P}_{22}\mathbf{P}_{33} - \\ &4\mathbf{P}_{14}^2 - 2\mathbf{P}_{11}\mathbf{P}_{44} - 4\mathbf{P}_{24}^2 - \\ &2\mathbf{P}_{22}\mathbf{P}_{44} - 4\mathbf{P}_{34}^2 - 2\mathbf{P}_{33}\mathbf{P}_{44}\} + \\ &\breve{\mathbf{H}}\mathbf{P}_0\breve{\mathbf{H}}^T + \breve{\mathbf{G}}\mathbf{V}\breve{\mathbf{G}}^T, \end{aligned} \tag{68}$$

where $\mathbf{P}_0$ has been partitioned as

$$\mathbf{P}_0 = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{P}_{13} & \mathbf{P}_{14} \\ \mathbf{P}_{12} & \mathbf{P}_{22} & \mathbf{P}_{23} & \mathbf{P}_{24} \\ \mathbf{P}_{13} & \mathbf{P}_{23} & \mathbf{P}_{33} & \mathbf{P}_{34} \\ \mathbf{P}_{14} & \mathbf{P}_{24} & \mathbf{P}_{34} & \mathbf{P}_{44} \end{bmatrix}. \tag{69}$$

7 EXPERIMENTAL RESULTS

The techniques developed in this paper were tested via Monte Carlo simulations. These simulations each used the same real GPS satellite ephemeris data collected beginning at 19:22:33.5 PST on Friday, October 13th 2000 at (-2.5192459e+06 m, -4.6431270e+06 m , 3.5626325e+06 m) in GPS earth centered earth fixed (ECEF) coordinates. Each simulation located the GPS receiver at a random position, centered at (-2.5192459e+06 m, -4.6431270e+06 m , 3.5626325e+06 m) in GPS ECEF coordinates, with standard deviation 1000 km. The noiseless pseudorange measurements were calculated, then corrupted with zero mean, 15 m standard deviation Gaussian noise. Thus, a sequence of artificial measurements where the true user position was known was available for testing our methods. A series of 50 simulations was performed, each simulation containing 1191 data points. The results of these simulations are displayed in Table 1.

The Monte Carlo simulations we used were fairly low fidelity, as they took no account of ionospheric or tropospheric noise. To check the validity of our results, we also ran a simulation on an Interstate

Table 1: Comparison of errors in Monte Carlo simulations

GPS solution method	mean error	error std. dev.
IDS	32.7163 m	1.3870 m
IDSBIS 2 steps	32.7143 m	1.3869 m
linear data only	581.0040 m	767.4165 m
project, then linearize	49.9955 m	69.8273 m
project, then min. var.	35.9738 m	20.7965 m
project, then max. lik.	32.7148 m	1.3866 m

Table 2: Comparison of errors from GPS satellite constellation simulator simulation

GPS solution method	mean error	error std. dev.
IDS	34.3087 m	0.9417 m
IDSBIS 2 steps	34.3196 m	0.9271 m
linear data only	50.4577 m	25.2115 m
project, then linearize	32.5180 m	2.0597 m
project, then min. var.	34.3197 m	0.9271 m
project, then max. lik.	34.3195 m	0.9271 m

Electronics Corporation model 2400 GPS satellite constellation simulator, collecting measurements with an Ashtech model Z-12 GPS receiver. This simulation followed the trajectory of an aircraft, initially located at $(962850.28547m, -5200816.32182m, 3563520.00371m)$ in GPS ECEF coordinates, starting at 16:30:00 PST on Monday, January 22, 2001. The corresponding measurement sequence, which consisted of 1680 measurement epochs, was thus corrupted by true receiver noise, as well as a good approximation of the tropospheric and ionospheric noises. As with the Gaussian Monte Carlo simulations, the true position of the antenna was known, allowing precise calculation of the estimation errors. The results of several solution techniques applied this simulation appear in Table 2.

The linear method alone is not as accurate as other methods, because all of the information in the measurements has not been used. When the methods of the last section are used, the results are of comparable accuracy to those of Biton, *et al.* In fact, the magnitude of the errors of the maximum likelihood method differ from those of the IDS only by millimeters.

The IDS scheme may be implemented in a recursive fashion, which is called the "IDS-Based Iterative Solution" (IDSBIS) in [6]. To our surprise, iterating the IDS failed to notably improve the accuracy of the solution in our Monte Carlo simulations, in contrast to the results reported in [6]. We speculate that this was due to the true Gaussian nature of the measurement errors in these simulations, whereas the real GPS measurements used by Biton *et al.* were corrupted by non-Gaussian noises.

8 CONCLUSIONS

This paper presents a direct method for solving the GPS equations. For noiseless pseudomeasurements, the user position can be determined by solving a set of linear equations, without making *any* approximations. If the pseudomeasurements are noisy, the equations are still linear in the unknown position and clock bias, and the nonlinearities in the noise terms are small enough to be safely ignored. The solutions are applicable to the differential GPS problem, as well as the single user GPS problem.

The conversion to a linear problem is wasteful in an information sense. That is, some of the measurement data is not present in the exact linear solution. The position estimate thus has a larger associated error covariance than that associated with an ILS method that has converged successfully. Of course, one cannot tell whether an ILS solution has converged to the correct answer, so a tradeoff has been made between certainty of convergence versus precision of the estimate. We have presented several methods for improving the linear estimate by using the information not present in the linear measurement equations. These techniques yield results on par with the ad hoc procedure developed by Biton *et al.* , while having a more sound theoretical basis and better understood error bounds and convergence guarantees.

ACKNOWLEDGMENTS

Our research was supported by the United States Air Force Office of Scientific Research under Grant F496200-00-0154, by NASA Dryden Flight Research Center under Grant NAS4-00051, and by NASA Goddard Space Flight Center under grant NAG5-8879.

REFERENCES

[1] P. S. Noe, K. A. Myers, and T. K. Wu, "A navigation algorithm for the low-cost GPS receiver," *Navigation. Journal of the Institute of Navigation*, vol. 25, pp. 258–64, Summer 1978.

[2] S. Bancroft, "An algebraic solution of the GPS equations," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-21, pp. 56–59, January 1985.

[3] L. O. Krause, "A direct solution to GPS-type navigation equations," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-23, pp. 225–32, March 1987.

[4] J. S. Abel and J. W. Chaffee, "Existence and uniqueness of GPS equations," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES–27, pp. 952–6, November 1991.

[5] J. W. Chaffee and J. S. Abel, "On the exact solutions of pseudorange equations," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES–30, pp. 1021–30, October 1994.

[6] I. Biton, M. Koifman, and I. Y. Bar-Itzhack, "Improved direct solution of the global positioning system equation," *AIAA Journal of Guidance, Control, and Dynamics*, vol. 21, pp. 45–49, January–February 1998.

[7] J. L. Leva, "An alternative closed-form solution to the GPS pseudo-range equations," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 32, pp. 1430–9, October 1996.

[8] C. Kee, "Wide area differential GPS," in *Global Positioning System: Theory and Applications* (B. W. Parkinson and J. J. Spilker Jr., eds.), vol. II, ch. 3, Washington, DC: American Institute of Aeronautics and Astronautics, Inc., 1996.

[9] D. G. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.

JOURNAL OF GUIDANCE, CONTROL, AND DYNAMICS
Vol. 25, No. 6, November–December 2002

Target Association Using Detection Methods

Jonathan D. Wolfe* and Jason L. Speyer†
University of California, Los Angeles, Los Angeles, California 90095-1597

A residual-based scheme is presented for solving the radar track-to-track association problem using bearings-only measurements. To accomplish track association between two stations, the residuals of a bank of nonlinear filters called modified gain extended Kalman filters are analyzed. Once tracks have been associated between two stations, tracks from additional stations may be associated with tracks from the first two stations by checking algebraic parity equations. Traditional track association methods rely on the local stations' estimated target positions and error variances. These local estimates may be quite inaccurate or even divergent when using bearings-only measurements. Our method bypasses this difficulty because our filters use raw data from multiple stations. An example demonstrates that our methods yield results superior to those of standard methods.

I. Introduction

SUPPOSE that several spatially distributed radar installations are each tracking several targets. Associating a given target to its track at each of the radar stations is an important issue, which the radar literature refers to as the track-to-track association problem. Suppose further that the stations use passive sensors that only measure bearings to the target, without measuring range. In this paper, we outline a strategy for solving this association problem by analyzing measurement residuals.

Bearings-only observation functions fall into two special classes of nonlinear functions, called modifiable and approximately modifiable nonlinearities, which are defined as follows:

Definition 1. A time-varying function $f:\mathbb{R}^n \to \mathbb{R}^q$ is called modifiable if there exists an operator $A:\mathbb{R}^q \times \mathbb{R}^n \to \mathbb{R}^{q\times n}$ such that for any $x, \bar{x} \in \mathbb{R}^n$,

$$f(x) - f(\bar{x}) = A[f(x), \bar{x}](x - \bar{x}) \tag{1}$$

Definition 2. A time-varying function $f:\mathbb{R}^n \to \mathbb{R}^q$ is called approximately modifiable if there exists a region $\mathcal{D} \subset \mathbb{R}^n$ and operators $A:\mathbb{R}^q \times \mathbb{R}^n \to \mathbb{R}^{q\times n}$ and $\mathcal{E}:\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{n\times n}$ such that for any x, $\bar{x} \in \mathcal{D}$,

$$f(x) - f(\bar{x}) = [A(f(x), \bar{x}) + \mathcal{E}(x, x - \bar{x})](x - \bar{x}) \tag{2}$$

where $\lim_{\|x-\bar{x}\|\to 0} \|\mathcal{E}(x, x-\bar{x})\|/\|A(f(x), \bar{x})\| = 0$.

Song and Speyer's modified gain extended Kalman filter (MGEKF)[1] is a globally convergent, unbiased, nonlinear observer for systems whose measurement functions are modifiable or approximately modifiable. In this paper, the observers we design for bearings-only track association are MGEKFs.

An early attempt at solving the track-to-track association problem was made by Singer and Kanyuck.[2] In their paper, they incorrectly assumed that estimation errors local to each station were uncorrelated. Bar-Shalom,[3] Bar-Shalom and Fortmann,[4] and Bar-Shalom and Campo[5] later corrected this error by accounting for the correlation between the local estimation errors due to the common process noise of the target. Later researchers have integrated the problem of track association directly into the process of separating the measurements corresponding to actual targets from clutter.[6,7] In all of these references, it is assumed that both range and bearings were measured. In some of these references, the possibility of using a MGEKF to handle the situation of bearings-only measurements is mentioned, but none have a discussion of the details of such an implementation, in particular problems associated with the asymmetry of single station estimation errors. Estimates based on bearings-only measurements from a single station are especially uncertain along the line between the target and the receiver. This uncertainty is reduced when measurements from physically separated stations are used. Our method attempts to take advantage of this phenomenon by using estimates constructed from several stations' measurements.

The paper is organized as follows. We show in Sec. II that bearings-only measurement functions are modifiable. (Prior results only showed that they were approximately modifiable.[1]) We then demonstrate in Sec. III that incorrect associations between two radar stations can be interpreted as sensor faults, so that a bank of modified-gain fault detection filters can be used to determine the track associations. Section IV contains the main result, an algorithm for solving the bearings-only track association problem. The application of this algorithm to an example in Sec. V compares our approach to a conventional track association method. Section VI concludes the paper.

In the sequel, inertial Cartesian coordinates describe the motion of each target in three dimensions via the state vector

$$x^t = [X^t \quad Y^t \quad Z^t \quad \dot{X}^t \quad \dot{Y}^t \quad \dot{Z}^t \quad \ddot{X}^t \quad \ddot{Y}^t \quad \ddot{Z}^t]^T \tag{3}$$

and the dynamics of each target are assumed to be of the form

$$x^t(k+1) = A(k)x^t(k) + B(k)w^t(k) \tag{4}$$

Note that we include an acceleration state to model maneuvering target dynamics.

II. Modifiability of Bearings-Only Measurements

Song and Speyer[1] showed that the azimuth angle $az_s^t \in [-\pi/2, \pi/2)$ and the elevation angle $el_s^t \in [-\pi/2, \pi/2)$ from station s to target t, as shown in Fig. 1, are modifiable and approximately modifiable, respectively. The region $\mathcal{D}$ in which the elevation angle was approximately modifiable excluded an ellipsoidal region near the sensor, making their algorithms difficult to implement for situations where the angular sensor gets close to the target, for example, in the terminal guidance of a missile. We improve this situation somewhat by introducing the new angle $\Psi_s^t \in [-\pi/2, \pi/2)$ and describing the position of the target in terms of Ψ_s^t and $\Phi_s^t \triangleq az_s^t$. Note that Ψ_s^t can be calculated from az_s^t and el_s^t via the equation

$$\Psi_s^t = \tan^{-1}\left(\frac{Z_s^t}{X_s^t}\right) = \tan^{-1}\left(\frac{\tan el_s^t}{\cos az_s^t}\right) \tag{5}$$

This section is devoted to proving that the measurement function for Ψ_s^t is modifiable.

Let $\bar{x}^t$ be an estimate of x^t and assume that the position of the measurement station in inertial space, $x_s = [X_s \quad Y_s \quad Z_s]$, is known.

Received 4 October 2001; revision received 1 May 2002; accepted for publication 12 June 2002.

*Research Engineer, Department of Mechanical and Aerospace Engineering.
†Professor, Department of Mechanical and Aerospace Engineering. Fellow AIAA.

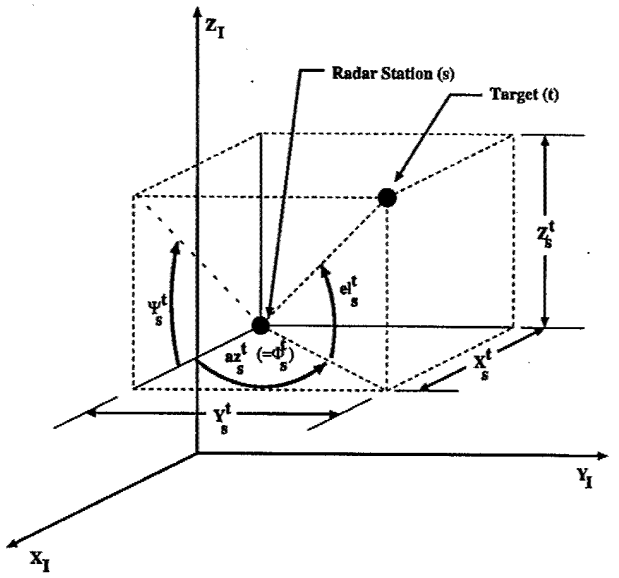


Fig. 1 Angles for target bearings.

Then X_s^t, Y_s^t, Z_s^t, $\bar{X}_s^t$, $\bar{Y}_s^t$, and $\bar{Z}_s^t$ can be computed by taking the difference between elements of x^t, $\bar{x}^t$, and x_s.

Suppose that station s measures the bearings of target t with the measurement vector z_s^t. Define $h_s(x^t)$ by

$$h_s(x^t) \triangleq \begin{bmatrix} \Phi_s^t \\ \Psi_s^t \end{bmatrix} = z_s^t \tag{6}$$

The measurement residual corresponding to $h_s(x^t)$ is then

$$h_s(x^t) - h_s(\bar{x}^t) = \begin{bmatrix} \tan^{-1}\left(Y_s^t/X_s^t\right) - \tan^{-1}\left(\bar{Y}_s^t/\bar{X}_s^t\right) \\ \tan^{-1}\left(Z_s^t/X_s^t\right) - \tan^{-1}\left(\bar{Z}_s^t/\bar{X}_s^t\right) \end{bmatrix} \triangleq \begin{bmatrix} \tan^{-1}\alpha \\ \tan^{-1}\beta \end{bmatrix} \tag{7}$$

Applying the trigonometric identity

$$\tan^{-1}(a) - \tan^{-1}(b) = \tan^{-1}[(a-b)/(1+ab)]$$

we obtain

$$\begin{bmatrix} \tan^{-1}\alpha \\ \tan^{-1}\beta \end{bmatrix} = \begin{bmatrix} \tan^{-1}\dfrac{\left(Y_s^t/X_s^t\right) - \left(\bar{Y}_s^t/\bar{X}_s^t\right)}{1 + \left(Y_s^t/X_s^t\right)\left(\bar{Y}_s^t/\bar{X}_s^t\right)} \\ \tan^{-1}\dfrac{\left(Z_s^t/X_s^t\right) - \left(\bar{Z}_s^t/\bar{X}_s^t\right)}{1 + \left(Z_s^t/X_s^t\right)\left(\bar{Z}_s^t/\bar{X}_s^t\right)} \end{bmatrix} \tag{8}$$

$$\begin{bmatrix} \tan^{-1}\alpha \\ \tan^{-1}\beta \end{bmatrix} = \begin{bmatrix} \tan^{-1}\left(\dfrac{Y_s^t\bar{X}_s^t - \bar{Y}_s^t X_s^t}{X_s^t\bar{X}_s^t + Y_s^t\bar{Y}_s^t}\right) \\ \tan^{-1}\left(\dfrac{Z_s^t\bar{X}_s^t - \bar{Z}_s^t X_s^t}{X_s^t\bar{X}_s^t + Z_s^t\bar{Z}_s^t}\right) \end{bmatrix} \tag{9}$$

Define

$$H\left(z_s^t\right) \triangleq \begin{bmatrix} \sin\left(\Phi_s^t\right) & -\cos\left(\Phi_s^t\right) & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \sin\left(\Psi_s^t\right) & 0 & -\cos\left(\Psi_s^t\right) & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{10}$$

Let $d_1 \triangleq \sqrt{[(X_s^t)^2 + (Y_s^t)^2]}$, $D_1 \triangleq d_1/[X_s^t\bar{X}_s^t + Y_s^t\bar{Y}_s^t]$, $d_2 \triangleq \sqrt{[(X_s^t)^2 + (Z_s^t)^2]}$, and $D_2 \triangleq d_2/[X_s^t\bar{X}_s^t + Z_s^t\bar{Z}_s^t]$. Note also that $\sin(\Phi_s^t) = Y_s^t/d_1$, $\cos(\Phi_s^t) = X_s^t/d_1$, $\sin(\Psi_s^t) = Z_s^t/d_2$, and $\cos(\Psi_s^t) = X_s^t/d_2$. Therefore, we can express D_1 and D_2 as functions of the estimates and measured angles:

$$D_1 = D_1\left(z_s^t, \bar{x}^t\right) = 1/\left[\cos\left(\Phi_s^t\right)\bar{X}_s^t + \sin\left(\Phi_s^t\right)\bar{Y}_s^t\right]$$

$$D_2 = D_2\left(z_s^t, \bar{x}^t\right) = 1/\left[\cos\left(\Psi_s^t\right)\bar{X}_s^t + \sin\left(\Psi_s^t\right)\bar{Z}_s^t\right]$$

If we express the trigonometric functions in $H(z_s^t)$, D_1, and D_2 in terms of X_s^t, Y_s^t, Z_s^t, $\bar{X}_s^t$, $\bar{Y}_s^t$, and $\bar{Z}_s^t$, we can write Eq. (9) as a function of z_s^t and $\bar{x}^t$:

$$\begin{bmatrix} \alpha\left(z_s^t, \bar{x}^t\right) \\ \beta\left(z_s^t, \bar{x}^t\right) \end{bmatrix} = \begin{bmatrix} D_1\left(z_s^t, \bar{x}^t\right) & 0 \\ 0 & D_2\left(z_s^t, \bar{x}^t\right) \end{bmatrix} H\left(z_s^t\right)\left[\bar{x}^t - x_s\right] \tag{11}$$

Finally, we can rewrite Eq. (11) as

$$\begin{bmatrix} -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1/\alpha\left(z_s^t, \bar{x}^t\right) & 0 \\ 0 & 1/\beta\left(z_s^t, \bar{x}^t\right) \end{bmatrix} \begin{bmatrix} D_1\left(z_s^t, \bar{x}^t\right) & 0 \\ 0 & D_2\left(z_s^t, \bar{x}^t\right) \end{bmatrix} \times H\left(z_s^t\right)\left[x^t - x^t - \bar{x}^t + x_s\right] \tag{12}$$

and combine it with Eq. (7) to obtain $h_s(x^t)$ in modifiable form,

$$h_s(x^t) - h_s(\bar{x}^t) = \begin{bmatrix} -\dfrac{D_1\left(z_s^t, \bar{x}^t\right)\tan^{-1}\alpha\left(z_s^t, \bar{x}^t\right)}{\alpha\left(z_s^t, \bar{x}^t\right)} & 0 \\ 0 & -\dfrac{D_2\left(z_s^t, \bar{x}^t\right)\tan^{-1}\beta\left(z_s^t, \bar{x}^t\right)}{\beta\left(z_s^t, \bar{x}^t\right)} \end{bmatrix} \times H\left(z_s^t\right)[x^t - \bar{x}^t] \tag{13}$$

where we have made use of the identity

$$H\left(z_s^t\right)[x_s - x^t] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Thus, we have replaced the elevation angle el_s^t, from which Song and Speyer[1] produced an approximately modifiable function with a new angle Ψ_s^t. Like the azimuth angle Φ_s^t, angle Ψ_s^t leads to modifiable measurement functions.

III. Converting Incorrect Associations into Sensor Faults

Suppose that station s can view several targets, indexed by i, and measures the bearings of each target. Then each of these measurements z_s^i is generated by $h_s(x^i)$, as in Eq. (6). Now suppose that another station, using its local observations, generates a state estimate of one of the targets that station s views. This estimate $\bar{x}^j$ corresponds to x^j, the true state of the jth target at station s, but neither station knows the value of index j. Our goal is to determine which of the tracks at station s is the jth one, using only $\{z_s^i\}$, the measurements local to station s, and $\bar{x}^j$, the other station's state estimate of one of the targets.

To this end, let us form the following error residual between the estimate $\bar{x}^j$ and the measurement z_s^i, making use of the result from the preceding section:

$$z_s^i - h_s(\bar{x}^j) = h_s(x^i) - h_s(\bar{x}^j) = G\left(z_s^i, \bar{x}^j\right)(x^i - \bar{x}^j) \tag{14}$$

where from Eq. (13)

$$G\left(z_s^i, \bar{x}^j\right) = \begin{bmatrix} -\dfrac{D_1\left(z_s^i, \bar{x}^j\right)\tan^{-1}\alpha\left(z_s^i, \bar{x}^j\right)}{\alpha\left(z_s^i, \bar{x}^j\right)} & 0 \\ 0 & -\dfrac{D_2\left(z_s^i, \bar{x}^j\right)\tan^{-1}\beta\left(z_s^i, \bar{x}^j\right)}{\beta\left(z_s^i, \bar{x}^j\right)} \end{bmatrix} \times H\left(z_s^i\right) \tag{15}$$

By introducing a zero term into the measurement residual, we can rephrase it as

$$z_s^i - h_s(\bar{x}^j) = h_s(x^i) - h_s(\bar{x}^j) \tag{16}$$

$$z_s^i - h_s(\bar{x}^j) = G\left(z_s^i, \bar{x}^j\right)(x^i - \bar{x}^j) \tag{17}$$

$$z_s^i - h_s(\bar{x}^j) = G\left(z_s^i, \bar{x}^j\right)(x^i - \bar{x}^j + x^j - x^j) \tag{18}$$

$$z_s^i - h_s(\bar{x}^j) = G\left(z_s^i, \bar{x}^j\right)(x^j - \bar{x}^j) + G\left(z_s^i, \bar{x}^j\right)(x^i - x^j) \tag{19}$$

$$z_s^i - h_s(\bar{x}^j) = G\left(z_s^i, \bar{x}^j\right)(x^j - \bar{x}^j) + \mu_s^{ij} \tag{20}$$

where $\mu_s^{ij} \triangleq G(z_s^i, \bar{x}^j)(x^i - x^j)$ represents the difference between x^i and x^j as a sensor fault. If $i = j$, we have correctly guessed the association between measurement and estimate, and there is no fault ($\mu_s^{ij} = 0$). If $i \neq j$, then $\mu_s^{ij} \neq 0$, playing the role of a sensor fault in the residual.

IV. Algorithm for Track Association from Bearings-Only Measurements via Fault Detection Filters

Suppose that there are S radar stations, with known inertial coordinates, that make bearings-only measurements in three-space of T different targets. We assume that all measurements at each station have been grouped as tracks of each target visible at that station using conventional means.[4,8,9] In this section, we propose an algorithm for associating the tracks at all stations to their corresponding targets.

Assume that each measurement station s is located at known inertial coordinates (X_s, Y_s, Z_s). Let $\bar{x}^{1i}$ denote a fault detection filter's estimate of the target corresponding to the ith track at the first station. The bearings-only measurement function for the station s of the same target is thus

$$h_s(x^{1i}) \triangleq \begin{bmatrix} \tan^{-1}\left(\dfrac{Y^{1i} - Y_s}{X^{1i} - X_s}\right) \\ \tan^{-1}\left(\dfrac{Z^{1i} - Z_s}{X^{1i} - X_s}\right) \end{bmatrix}$$

From the results of the preceding section, the error residual of track j at any station s, generated by target i at station 1, is given by

$$z_s^j - h_s(\bar{x}^{1i}) \approx G\left(z_s^j, \bar{x}^{1i}\right)(x^{1i} - \bar{x}^{1i}) + \mu_s^{ij} + v^j \tag{21}$$

where $G(z_s^j, \bar{x}^{1i})$ is given by Eq. 15 and the sensor noise is

$$v^j = \mathcal{N}(0, V^j)$$

The approximate structure of Eq. (21) is due to the replacement of the measurement function in $G(\cdot, \cdot)$ with the actual measurement (see Song and Speyer[1]). Note that, by default, $\mu_1^{ii} = 0, \forall i = 1, \ldots, T$.

The following algorithm, illustrated in Fig. 2, associates tracks between stations.

Algorithm (track association):

1) Let $i = 1$.

2) Run a bank of T detection filters that operate on data from stations 1 and 2, where the jth filter attempts to detect μ_2^{ij}. Each filter is constructed using the dynamic detection filter procedure given next. All but one of these detection filters should register a fault. The track corresponding to the filter that detected no fault is associated with z_1^i. Without loss of generality, label this track z_2^i.

3) For each track z_s^l, $s = 3, \ldots, S$, $l = 1, \ldots, T$, perform the algebraic parity test given subsequently. If the result of the parity test is zero, then z_s^l is associated with z_1^i and z_2^i.

4) If $i < T$, increment i by 1 and go to step 2. If $i = T$, we have completed the track association procedure.

Note that estimates obtained in step 2 are used in step 3. Therefore, stations 1 and 2 should be chosen to maximize observability of the targets.

Dynamic Detection Filter

For any estimator of x^{1i}, the estimation residual determined by the measurements z_1^i and z_2^j will not converge to values near zero unless z_1^i and z_2^j correspond to the same target. One such estimator is the MGEKF[1] given as

$$\bar{x}^{1i}(k+1) = A(k)\hat{x}^{1i}(k) \tag{22}$$

$$r^{ij}(k) = \begin{Bmatrix} z_1^i(k) - h_1[\hat{x}^{1i}(k)] \\ z_2^j(k) - h_2[\hat{x}^{1i}(k)] \end{Bmatrix} \tag{23}$$

$$\hat{x}^{1i}(k) = \bar{x}^{1i}(k) + K^{ij}(k)r^{ij}(k) \tag{24}$$

$$M^{ij}(k+1) = A(k)P^{ij}(k)A^T(k) + Q(k) \tag{25}$$

$$\bar{h}_{x^{1i}(k)} = \begin{bmatrix} -\dfrac{\bar{Y}^{1i} - Y_1}{\left\{1 + \left[\left(\bar{Y}^{1i} - Y_1\right)/\left(\bar{X}^{1i} - X_1\right)\right]^2\right\}\left(\bar{X}^{1i} - X_1\right)^2} & \dfrac{1}{\left\{1 + \left[\left(\bar{Y}^{1i} - Y_1\right)/\left(\bar{X}^{1i} - X_1\right)\right]^2\right\}\left(\bar{X}^{1i} - X_1\right)^2} & \cdots \\ -\dfrac{\bar{Z}^{ij} - Z_1}{\left\{1 + \left[\left(\bar{Z}^{ij} - Z_1\right)/\left(\bar{X}^{1i} - X_1\right)\right]^2\right\}\left(\bar{X}^{1i} - X_1\right)^2} & 0 & \\ -\dfrac{\bar{Y}^{1i} - Y_2}{\left\{1 + \left[\left(\bar{Y}^{1i} - Y_2\right)/\left(\bar{X}^{1i} - X_2\right)\right]^2\right\}\left(\bar{X}^{1i} - X_2\right)^2} & \dfrac{1}{\left\{1 + \left[\left(\bar{Y}^{1i} - Y_2\right)/\left(\bar{X}^{1i} - X_2\right)\right]^2\right\}\left(\bar{X}^{1i} - X_2\right)^2} & \\ -\dfrac{\bar{Z}^{ij} - Z_2}{\left\{1 + \left[\left(\bar{Z}^{ij} - Z_2\right)/\left(\bar{X}^{1i} - X_2\right)\right]^2\right\}\left(\bar{X}^{1i} - X_2\right)^2} & 0 & \end{bmatrix}$$

$$\cdots \begin{bmatrix} 0 & 0\ 0\ 0\ 0\ 0\ 0 \\ \dfrac{1}{\left\{1 + \left[\left(\bar{Z}^{ij} - Z_1\right)/\left(\bar{X}^{1i} - X_1\right)\right]^2\right\}\left(\bar{X}^{1i} - X_1\right)^2} & 0\ 0\ 0\ 0\ 0\ 0 \\ 0 & 0\ 0\ 0\ 0\ 0\ 0 \\ \dfrac{1}{\left\{1 + \left[\left(\bar{Z}^{ij} - Z_2\right)/\left(\bar{X}^{1i} - X_2\right)\right]^2\right\}\left(\bar{X}^{1i} - X_2\right)^2} & 0\ 0\ 0\ 0\ 0\ 0 \end{bmatrix} \tag{26}$$

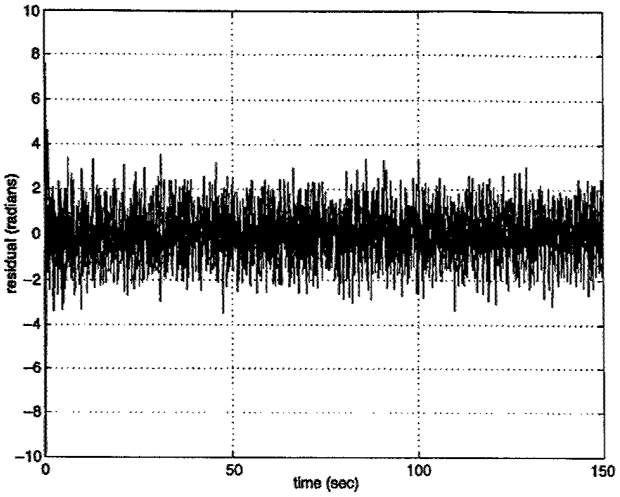


Fig. 3 MGEKF residual for matching tracks.

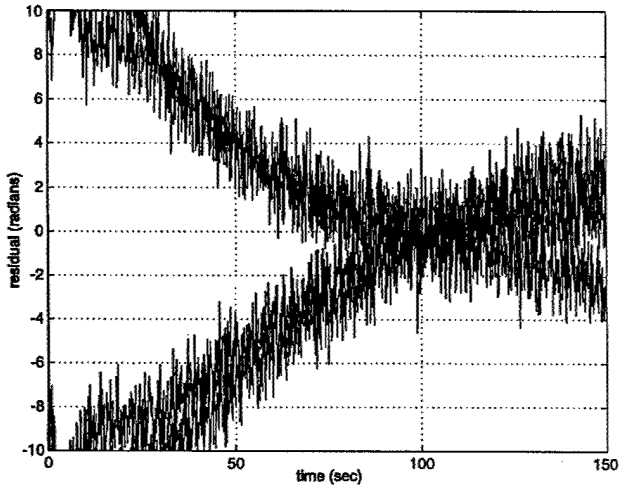


Fig. 4 MGEKF residual for mismatched tracks.

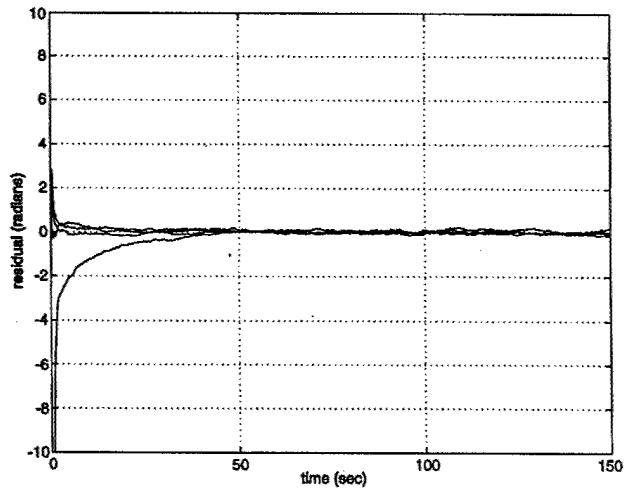


Fig. 5 Filtered MGEKF residual for matching tracks.

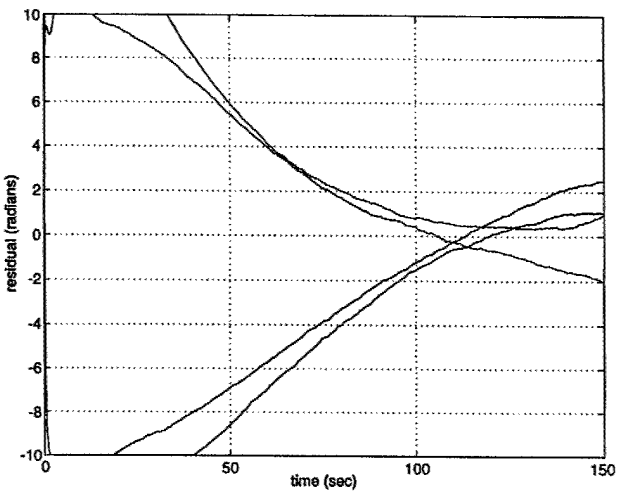


Fig. 6 Filtered MGEKF residual for mismatched tracks.

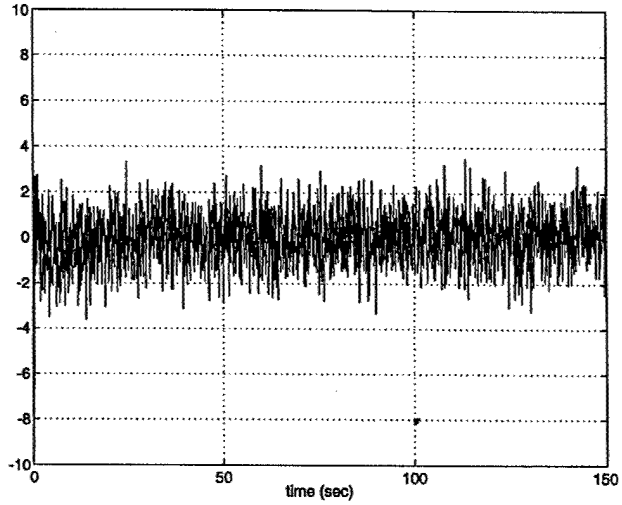


Fig. 7 Parity test residual for matching tracks.

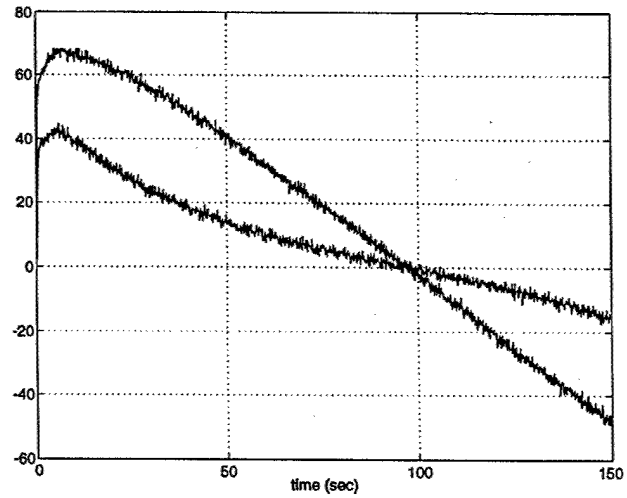


Fig. 8 Parity test residual for mismatched tracks.

covariance 10^{-3}, and measurement noise with covariance 1). These estimates (Figs. 5 and 6) clearly show that the mean corresponding to a mismatch looks nothing like that of the matched case.

After the tracks had been associated between the first two stations, algebraic parity tests attempted to associate the targets observed by the third station relative to those observed by the first and second stations. Two plots of residuals generated by the algebraic parity tests appear in Figs. 7 and 8. Again, the residuals for the mismatch are much larger than those corresponding to a correct association.

For purposes of comparison, Fig. 9 plots the error statistic developed by Bar-Shalom[3] and Bar-Shalom and Fortmann[4] for both a correct and an incorrect track association (using the same data sequences that were used by the filters in Figs. 3 and 4). Note that the chi-squared error statistic does not change much between the matched and mismatched cases. We also noticed that there

Table 1 Radar station positions

Station identification	x Position, m	y Position, m	z Position, m
1	50	1	0
2	50,000	1	50
3	25,000	−400	100

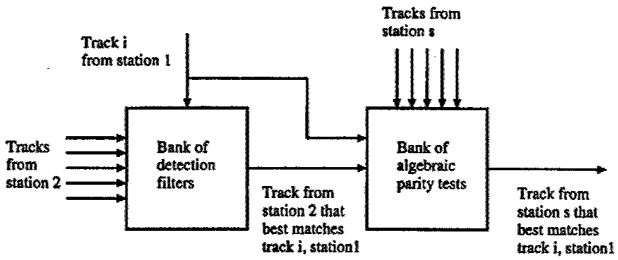


Fig. 2 Track association procedure.

$$K^{ij}(k) = M^{ij}(k)\tilde{h}^T_{x^{1i}(k)}\left[\tilde{h}_{x^{1i}(k)}M^{ij}(k)\tilde{h}^T_{x^{1i}(k)} + V^{ij}(k)\right]^{-1} \tag{27}$$

$$\tilde{G}\left[z_1^i(k), z_2^j(k), \bar{x}^{1i}(k)\right] = \begin{bmatrix} G\left(z_1^i(k), \bar{x}^{1i}(k)\right) \\ G\left(z_2^j(k), \bar{x}^{1i}(k)\right) \end{bmatrix} \tag{28}$$

$$P^{ij}(k) = \left\{I - K^{ij}(k)\tilde{G}\left[z_1^i(k), z_2^j(k), \bar{x}^{1i}(k)\right]\right\}M^{ij}(k)$$
$$\times \left\{I - K^{ij}(k)\tilde{G}\left[z_1^i(k), z_2^j(k), \bar{x}^{1i}(k)\right]\right\}^T$$
$$+ K^{ij}(k)(V^{ij})^{-1}(k)(K^{ij})^T(k) \tag{29}$$

where

$$V^{ij}(k) = \text{diag}\{V^i, V^j\} \tag{30}$$

The weighted innovations process of the MGEKF,

$$\nu^{ij}(k) = \left[\tilde{h}_{x^{1i}(k)}M^{ij}(k)\tilde{h}^T_{x^{1i}(k)} + V^{ij}(k)\right]^{\frac{1}{2}} r^{ij}(k) \tag{31}$$

should be close to a zero-mean, unit variance white noise sequence only if z_1^i and z_2^j correspond to the same target.

Algebraic Parity Test

This test determines if z_s^l, $S \geq s > 2$, $T \geq l \geq 1$, is associated with z_1^i and z_2^j, where z_1^i and z_2^j are already known to be associated with each other. Suppose that $\hat{x}^{1i}$ is the state estimate generated by z_1^i and z_2^j. Then, if z_s^l is associated with the tracks z_1^i and z_2^j,

$$\nu(k)_l^{1i} \triangleq \left[\tilde{h}_{x^{1i}(k)}M^{ij}(k)\tilde{h}^T_{x^{1i}(k)} + V^{ij}(k)\right]^{\frac{1}{2}}\left\{z_s^l(k) - h_s[\hat{x}^{1i}(k)]\right\} \tag{32}$$

should be close to a zero mean, unit variance white noise sequence. Here, the approximate measurement matrix $\tilde{h}_{x^{1i}(k)}$ is computed in a manner similar to the first two rows of the matrix in Eq. (26), but referenced to (X_s, Y_s, Z_s), the location of station s, instead of the location of the first station (X_1, Y_1, Z_1). The algebraic parity test is simply to evaluate the parity equation (32).

V. Example

The track association algorithm presented in the last section is applied to simulation data in this section. Three radar installations were located at the positions given by Table 1, and two targets were both modeled as ninth-order linear time-invariant discrete-time systems with the dynamics

$$\dot{x}(t) = Fx(t) + \Gamma w(t) \tag{33}$$

where

$$F \triangleq \begin{bmatrix} 0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0 \\ 0&0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&-\alpha&0&0 \\ 0&0&0&0&0&0&0&-\alpha&0 \\ 0&0&0&0&0&0&0&0&-\alpha \end{bmatrix}$$

$$\Gamma \triangleq \begin{bmatrix} 0&0&0 \\ 0&0&0 \\ 0&0&0 \\ 0&0&0 \\ 0&0&0 \\ 0&0&0 \\ 1&0&0 \\ 0&1&0 \\ 0&0&1 \end{bmatrix} \tag{34}$$

and where w is a zero mean Brownian motion process with covariance $I_{3\times 3}$ and $\alpha = \frac{1}{10}$ is the time constant for the first-order filters that model target maneuvers as colored noise processes. We sample this model at intervals of $T = 0.1$ s to generate the discrete time dynamics

$$x(k+1) = Ax(k) + Bw(k) \tag{35}$$

where

$$A = e^{FT}, \qquad B = \int_0^T e^{Ft}B\,dt, \qquad E[w(k)] = 0_{3\times 1}$$
$$E[w(k)w^T(l)] = I_{3\times 3}\delta_{kl} \tag{36}$$

The targets began the simulation with the initial conditions

$$x_1(0) = [50 \quad 220{,}000 \quad 30{,}000 \quad 250 \quad -1000 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

$$x_2(0) = [50{,}000 \quad 20{,}000 \quad 35{,}000 \quad -250 \quad 1000 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

This configuration corresponds to the two targets initially moving directly toward each other, in a line that almost passes through station 2. In the simulation, they pass closest to each other at $t = 99.2$ s. Each measurement station measures the angles Φ_s^t and Ψ_s^t to each target at every sample time. These measurements are subject to additive, normally distributed zero-mean white measurement noise with standard deviation 1 deg. We assume that the measurement noise is independent between sensors at all stations. Each MGEKF begins with the a priori information

$$\hat{x}^{1i}(0) = [25{,}000 \quad 120{,}000 \quad 32{,}500 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

$$P^{ij}(0) = 10^7 \times I_{9\times 9}$$

Finally, we assume that the local stations were able to separate their measurements from clutter perfectly using methods like those of Reid[9] or Bar-Shalom and Fortmann,[4] or Fortmann and Bar-Shalom.[8]

Figure 3 plots the weighted innovations of a MGEKF that uses measurements from stations 1 and 2 that correspond to the second target, whereas Fig. 4 plots the weighted innovations of a MGEKF that uses measurements that are mismatched. Note that the innovations for the correct match appear to be a zero mean white noise sequence, whereas the innovations for the incorrect match are larger and are not white. To better observe the behavior of these sequences, their means were estimated using a Kalman filter (assuming that each element of the weighted innovation of the MGEKF was a measurement of a process that had integrator dynamics, process noise with

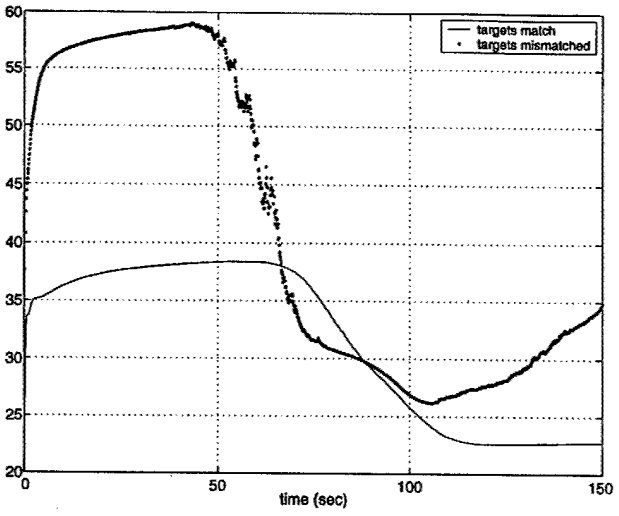


Fig. 9 Error statistic suggested by Bar-Shalom[3] and Bar-Shalom and Fortmann[4]: $(\hat{x}_1 - \hat{x}_2)^T E[(\hat{x}_1 - \hat{x}_2)(\hat{x}_1 - \hat{x}_2)^T](\hat{x}_1 - \hat{x}_2)$.

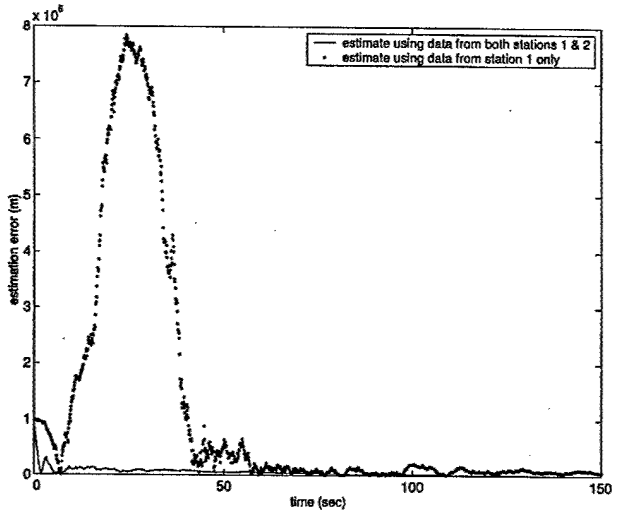


Fig. 10 Euclidean norm of error in tracking target 1.

were several instances where nearly singular matrices were inverted in the algorithm that computes the covariance of the difference between two local estimates.

Part of the reason for this difficulty is explained in Fig. 10, a plot of the Euclidean norm of the estimation error. The solid line corresponds to a MGEKF that uses measurements from both station 1 and 2, whereas the dotted line is from a filter that only used station 1 measurements. Any method that relies on estimates that only use a single station's measurements is subject to a large error. This is not a huge concern for linear estimators, but the matrix P^{ij} defined by Eq. (29) may not necessarily reflect this error.

We have also encountered cases where a single station measurement MGEKF was divergent in the radial direction to the target, but no such difficulties have appeared when data from two geographically disparate stations was used. One way of generating such a divergent case was to decrease the maneuver colored noise autocorrelation parameter α to $\frac{1}{20}$ or below. We note that values of this parameter below $\frac{1}{20}$ correspond to slower maneuvers, a commonly encountered situation.

VI. Conclusions

This paper describes residual-based techniques for solving the radar track association problem for bearings-only measurements. The association between the tracks at two stations can be determined by examining the residuals of a bank of MGEKFs. Once this association is established, an algebraic parity test can find the correspondence between tracks at other stations and targets tracked by the first two stations.

One may ask why detection filters are necessary: Why not do everything with algebraic parity tests? Although the detection filtering step is not strictly necessary, it does improve the quality of the track associations because the state estimates constructed from two widely separated stations are so much more accurate than the estimates from a single station.

To ensure the quality of the estimates from the MGEKFs, one could delay the algebraic parity testing steps for associating tracks from additional stations. If these parity tests are replaced with additional detection filter banks until the estimates before and after including a new station's measurements are sufficiently close, then the fidelity of the estimates can be guaranteed.

Acknowledgments

This research was supported in part by the Air Force Office of Scientific Research under Grant F49620-00-1-0154 and by Sandia Laboratory under U.S. Department of Energy Grant LH-1376.

References

[1]Song, T. L., and Speyer, J. L., "A Stochastic Analysis of a Modified Gain Extended Kalman Filter with Applications to Estimation with Bearings Only Measurements," *IEEE Transactions on Automatic Control*, Vol. AC-30, No. 10, 1985, pp. 940–949.

[2]Singer, R. A., and Kanyuck, A. J., "Computer Control of Multiple Site Track Correlation," *Automatica*, Vol. 7, No. 4, 1971, pp. 455–463.

[3]Bar-Shalom, Y., "On the Track-to-Track Correlation Problem," *IEEE Transactions on Automatic Control*, Vol. AC-26, No. 2, 1981, pp. 571, 572.

[4]Bar-Shalom, Y., and Fortmann, T. E., *Tracking and Data Association*, Vol. 179, Mathematics in Science and Engineering, Academic Press, Boston, 1988, pp. 217–265, 266–272.

[5]Bar-Shalom, Y., and Campo, L., "The Effect of the Common Process Noise on the Two-Sensor Fused-Track Covariance," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-22, No. 6, 1986, pp. 803–805.

[6]Pao, L. Y., "Multisensor Multitarget Mixture Reduction Algorithms for Tracking," *Journal of Guidance, Control, and Dynamics*, Vol. 17, No. 6, 1994, pp. 1205–1211.

[7]Pao, L. Y., "Measurement Reconstruction Approach for Distributed Multisensor Fusion," *Journal of Guidance, Control, and Dynamics*, Vol. 19, No. 4, 1996, pp. 842–847.

[8]Fortmann, T. E., and Bar-Shalom, Y., "Sonar Tracking of Multiple Targets Using Joint Probabalistic Data Association," *Journal of Oceanic Engineering*, Vol. OE-8, No. 3, 1983, pp. 173–183.

[9]Reid, D. B., "An Algorithm for Tracking Multiple Targets," *IEEE Transactions on Automatic Control*, Vol. AC-24, No. 6, 1979, pp. 843–854.

Appendix L

"Multiple Hypothesis Sequential Probability Ratio Tests for Resolving Integer Ambiguity in GPS,"

Jonathan D. Wolfe, Jason L. Speyer, and Walton R. Williamson,

Proceedings of the ION GPS 2001 and to be published in *The Journal of the Institute of Navigation*

Hypothesis Testing for Resolving Integer Ambiguity in GPS

Jonathan D. Wolfe *University of California, Los Angeles*
Walton R. Williamson *University of California, Los Angeles*
Jason L. Speyer *University of California, Los Angeles*

BIOGRAPHIES

Jonathan D. Wolfe is a Ph.D. candidate in the Mechanical and Aerospace Engineering Department at the University of California, Los Angeles.

Walton R. Williamson is a research engineer in the Mechanical and Aerospace Engineering Department at the University of California, Los Angeles. He received his Ph.D. degree in mechanical engineering from UCLA in December, 2000.

Jason L. Speyer is a professor in the Mechanical and Aerospace Engineering Department at the University of California, Los Angeles. He received his Ph.D. degree in applied mathematics at Harvard University, Cambridge. Prof. Speyer is a Fellow of both the American Institute of Aeronautics and Astronautics and the Institute of Electrical and Electronics Engineers.

ABSTRACT

In this paper, we present two statistical techniques appropriate for the "validation" of integer ambiguities and the detection of cycle slips. The multiple hypothesis Wald sequential probability ratio test (SPRT) can find the conditional probability that each set of integer biases under consideration is the true bias condition. The multiple hypothesis Shiryayev SPRT determines the conditional probability that the integer biases have jumped from the nominal bias condition to each member of a collection of other bias conditions. Hence, the Wald SPRT is a method for validating the integer ambiguities during the initial ambiguity resolution, while the Shiryayev SPRT can be used to monitor for cycle slips.

Each of these multiple hypothesis SPRTs (MHSPRTs) makes use of two measurement residuals. One is geometric combination of the carrier phase measurements, and the other is generated by differencing the carrier phase measurements with code measurements.

Prior work on cycle slip monitoring has focused solely on the detection of the occurrence of a cycle slip in the fastest time, balanced against the probability of issuing a false alarm. Once a disruption has occurred, the ambiguity resolution process must restart from scratch. The Shiryayev SPRT bypasses this problem, as it announces the location of the biases after the jump, in addition to the time of the cycle slip.

The calculations for the MHSPRTs are not linked to any particular distribution, unlike prior efforts. Only the probability density functions of the measurement residuals are required. Hence, the techniques can correctly compensate for non-Gaussian errors in measurement such as multipath.

For each hypothesis under consideration, the MHSPRTs yield the probability of that hypothesis being the correct one. The "state" of the MHSPRT recursions is the vector of all of these probabilities. Information from past measurements is embedded in this state. This recursive, probabilistic framework makes it very straightforward to add new hypotheses into the set of possible bias conditions while retaining information from prior measurements.

Results from successful simulations and field experiments are presented, showing the efficacy of our techniques.

1 INTRODUCTION

This paper describes new techniques for resolving a particular problem inherent in determining relative positions using the Global Positioning System (GPS). GPS was originally designed to determine the positions of antennae relative to the Earth, but when one is interested only in the positions of two antennae relative to each other, more precise "differential GPS" (DGPS) methods may be used. The most precise DGPS method is carrier phase DGPS, which measures the difference in the phase of the GPS carrier signal between two receivers. To create a relative position estimate using carrier phase GPS, the unknown number of full cycles of the carrier signal (the "integer

ambiguity") between the GPS receivers must be found and added to the differential phase. Standard least-squares estimation techniques generate floating point estimates of the integer ambiguity that can narrow the space of cycle numbers that must be searched in order to calculate ranges accurately. The small number of unknown cycles that could correspond to the error in the floating point estimate comprise the set of biases from which the integer ambiguity resolution algorithm must choose. This paper proposes several new algorithms for integer ambiguity resolution and cycle slip detection based on two statistical tests – the multiple hypothesis Wald probability ratio test and the multiple hypothesis Shiryayev probability ratio test.

There are several methods currently in use for resolving the integers, of which Teunissen's least-squares ambiguity decorrelation adjustment (LAMBDA) method [1, 2] is the most popular. The other commonly used integer resolution method is the ambiguity function method of Counselman and Gourevitch, and its' variants [3, 4, 5]. Both of these methods generate estimates of the differential position in the process of determining the integers. In contrast, the method recently developed by Park, *et al.* [6, 7] eliminates the differential state from the residual they use to determine the integers. We will use this residual in our development, since it's independence in time makes it valuable for statistical tests.

All of these prior methods use either Chi-squared tests or F-tests, so they all can potentially benefit from the more sophisticated recursive statistical methods described in this paper.

Mertikas and Rizos [8] have developed a scheme for detecting cycle slips in carrier phase GPS measurements. In their paper, they apply the CUSUM test [9, 10] to the residual of a Kalman filter in order to detect cycle slips. In order for this scheme to work, they must assume a dynamical structure for the integers that is somewhat artificial. Also, while their method will announce when a cycle skip has occurred, it cannot determine the position of the new integer bias. Once a cycle slip has been detected by their methods, a new integer ambiguity search must begin.

In contrast, the cycle slip detection methods we propose in the next sections explicitly announce the new integer ambiguities, as well as the time of the cycle slip. The only assumption about cycle slips that we make is an *a priori* probability of a slip, which is far less of a leap than constructing dynamics for the disruption. The statistical tests we use are also more computationally efficient than CUSUM. As with our methods for determining the initial integer ambiguities, the residual we analyze requires no estimation of the relative position between the GPS antennae.

The paper is summarized as follows: The Shiryayev and Wald multiple hypothesis sequential probability ratio tests are derived in Section 2. The improved integer resolution algorithm is presented in Section 3. Section 4 presents a residual that enables the designer to increase computational efficiency in exchange for longer sampling periods. A similar residual, presented in Section 5, allows the integer ambiguity associated with a newly acquired satellite to be rapidly determined when the other integer ambiguities are already known. Section 6 contains experimental results using data from simulations and from actual GPS measurements. The paper concludes with Section 7.

2 MULTIPLE HYPOTHESIS SEQUENTIAL PROBABILITY RATIO TEST

In this section we present two sophisticated statistical tests for determining the most likely event from a set of hypotheses. The multiple hypothesis Shiryayev sequential probability ratio test (MHSSPRT) detects jumps from a base hypothesis to another hypothesis in the set. The multiple hypothesis Wald sequential probability ratio test (MHWSPRT), which is a special case of the MHSSPRT, determines the most likely event from a set of hypotheses, assuming that the event is true for all time.

The MHSSPRT and MHWSPRT can be applied in place of the Chi-squared test in [6, 7]. The MHWSPRT yields somewhat better convergence times than Chi-squared, and the MHSSPRT allows one to monitor for cycle slips. However, the best improvement comes when these tests are applied to the enlarged residual presented in the next section.

The material in this section is adapted from the work of Malladi and Speyer [11].

2.1 RECURSIVE RELATION FOR SHIRYAYEV SEQUENTIAL PROBABILITY RATIO TEST

Suppose that we have a set of different hypotheses $\{\mathcal{H}_0, \mathcal{H}_1, \mathcal{H}_2, \ldots \mathcal{H}_m\}$. We wish to know if there is a transition from the base hypothesis $\mathcal{H}_0$ to any of the other hypotheses, and the time that the transition occurs. We will derive a recursive formula that at each time step computes the probability that a transition has occurred to each hypothesis, given the measurement residual sequence up to that time.

Let us define the following notation in this section:

$\mathbf{r}(k)$	Measurement residual vector at time k.
$\mathbf{R}(k)$	Measurement residual history up to time k.
θ_i	Time of transition to hypothesis $\mathcal{H}_i$.
$\mathcal{E}_i(k)$	Event $\{\theta_i \leq k+1\}$.
$F_i(k)$	$P(\theta_i \leq k \vert \mathbf{R}(k))$.
π_i	$P(\theta_i \leq 0)$.
$\tilde{p}_i$	*A priori* probability of transition to hypothesis $\mathcal{H}_i$ from time k to $k+1$.
$f_i(\cdot)$	Probability density function of $\mathbf{r}$ given $\mathcal{H}_i$.
$f_0(\cdot)$	Probability density function of $\mathbf{r}$ given $\mathcal{H}_0$.
$m+1$	Number of hypotheses.
$\phi_i(k)$	$P(\theta_i \leq k+1 \vert \mathbf{R}(k))$.

We assume that the measurement residual sequence $\{\mathbf{r}(k)\}$ is conditionally independent, *i.e.* the measurement residual sequence is independent once a disruption occurs. We also assume that the probability distributions of $\mathbf{r}(k)$ given $\mathcal{H}_i$ are known for every i. In particular, all of the probability density functions $f_i(\cdot)$ are known.

We will derive a recursive relation for $F_i(k)$, $i = 0, 1, 2, \ldots, m$.

Note first that there is a simple relation between $\phi_i(k)$ and $F_i(k)$ for $i > 0$:

$$\begin{aligned}
\phi_i(k) &= P(\theta_i \leq k+1 \vert \mathbf{R}(k)) && (1)\\
&= P(\theta_i \leq k \vert \mathbf{R}(k)) + P(\theta_i = k+1 \vert \mathbf{R}(k)) && (2)\\
&= P(\theta_i \leq k \vert \mathbf{R}(k)) + \\
&\quad P(\theta_i = k+1 \vert \theta_i > k, \mathbf{R}(k)) \cdot \\
&\quad\quad P(\theta_i > k \vert \mathbf{R}(k)) && (3)\\
&= F_i(k) + \tilde{p}_i \cdot (1 - F_i(k)). && (4)
\end{aligned}$$

Computing $\phi_0(k)$ is slightly more complicated. Define the set of events $\{\bar{\mathcal{E}}_i(k)\}$ so that $\bar{\mathcal{E}}_i(k)$ is the complement of $\mathcal{E}_i(k)$ for $i = 1, 2, \ldots, m$. If we assume that the events $\{\bar{\mathcal{E}}_i(k)\}$ are independent of each other, the probability of no transition before time $k+1$ is given by

$$\begin{aligned}
\phi_0(k) &= 1 - P\Big(\bigcup_{i=1}^{m} \mathcal{E}_i(k) \Big| \mathbf{R}(k)\Big) && (5)\\
&= P\Big(\bigcap_{i=1}^{m} \bar{\mathcal{E}}_i(k) \Big| \mathbf{R}(k)\Big) && (6)\\
&= \prod_{i=1}^{m} P(\bar{\mathcal{E}}_i(k) \vert \mathbf{R}(k)) && (7)\\
&= \prod_{i=1}^{m} \{1 - P(\mathcal{E}_i(k) \vert \mathbf{R}(k))\} && (8)\\
&= \prod_{i=1}^{m} \{1 - P(\theta_i \leq k+1 \vert \mathbf{R}(k))\} && (9)\\
&= \prod_{i=1}^{m} \{1 - \phi_i(k)\} && (10)
\end{aligned}$$

Lemma 2.1. *$F_i(k+1)$ is a function of $F_j(k)$, $j = 0, 1, \ldots, m$ in the following manner:*

$$F_i(k+1) = \frac{\phi_i(k) \cdot f_i(\mathbf{r}(k+1))}{\sum_{j=0}^{m} \phi_j(k) \cdot f_j(\mathbf{r}(k+1))}. \qquad (11)$$

Proof. By induction. We begin by showing that

$$F_i(1) = \frac{\phi_i(0) \cdot f_i(\mathbf{r}(1))}{\sum_{j=0}^{m} \phi_j(0) \cdot f_j(\mathbf{r}(1))}. \qquad (12)$$

By Bayes' Rule,

$$\begin{aligned}
F_i(1) = P(\theta_i \leq 1 \vert \mathbf{r}(1)) &= \frac{P(\mathbf{r}(1), \theta_i \leq 1)}{P(\mathbf{r}(1))} \\
&= \frac{P(\mathbf{r}(1) \vert \theta_i \leq 1) \cdot P(\theta_i \leq 1)}{P(\mathbf{r}(1))} && (13)
\end{aligned}$$

$$P(y(1)) = \sum_{j=0}^{m} P(\mathbf{r}(1) \vert \theta_j \leq 1) \cdot P(\theta_j \leq 1) \qquad (14)$$

$$\begin{aligned}
\phi_j(0) &= P(\theta_j \leq 1 \vert \text{no measurements}) \\
&= P(\theta_j \leq 1) \qquad \forall j. && (15)
\end{aligned}$$

Also,

$$P(y(1) \vert \theta_j \leq 1) = f_j(\mathbf{r}(1)) \cdot d\mathbf{r}(1) \qquad \forall j. \qquad (16)$$

Hence,

$$\begin{aligned}
F_i(1) = \frac{P(\mathbf{r}(1) \vert \theta_i \leq 1) \cdot P(\theta_i \leq 1)}{P(\mathbf{r}(1))} &= \\
\frac{f_i(\mathbf{r}(1)) \cdot d\mathbf{r}(1) \cdot \phi_i(0)}{\sum_{j=0}^{m} f_j(\mathbf{r}(1)) \cdot d\mathbf{r}(1) \cdot \phi_j(0)} &= \\
\frac{f_i(\mathbf{r}(1)) \cdot \phi_i(0)}{\sum_{j=0}^{m} f_j(\mathbf{r}(1)) \cdot \phi_j(0)}. && (17)
\end{aligned}$$

We next show that if we know $\{\mathbf{F}_0(k), \mathbf{F}_1(k), \ldots \mathbf{F}_m(k)\}$, then

$$F_i(k+1) = \frac{\phi_i(k) \cdot f_i(\mathbf{r}(k+1))}{\sum_{j=0}^{m} \phi_j(k) \cdot f_j(\mathbf{r}(k+1))} \quad \forall k > 0, \qquad (18)$$

which is a function of $\{\mathbf{F}_0(k), \mathbf{F}_1(k), \ldots \mathbf{F}_m(k)\}$ via the relations (4) and (10). At stage $k+1$,

$$\begin{aligned}
F_i(k+1) &= P(\theta_i \leq k+1 \vert \mathbf{R}(k+1)) \\
&= \frac{P(\mathbf{R}(k+1) \vert \theta_i \leq k+1) \cdot P(\theta_i \leq k+1)}{P(\mathbf{R}(k+1))} && (19)
\end{aligned}$$

$$P(\mathbf{R}(k+1)) = P(\mathbf{r}(k+1) \vert \mathbf{R}(k)) \cdot P(\mathbf{R}(k)) \qquad (20)$$

$$P(\mathbf{R}(k) \vert \theta_i \leq k+1) = \frac{P(\theta_i \leq k+1 \vert \mathbf{R}(k)) \cdot P(\mathbf{R}(k))}{P(\theta_i \leq k+1)} \qquad (21)$$

$$P(\mathbf{r}(k+1)|\theta_j \leq k+1) = f_j(\mathbf{r}(k+1)) \cdot d\mathbf{r}(k+1) \quad \forall j. \quad (22)$$

We now use the conditional independence of $\{\mathbf{r}(k)\}$ to write

$$F_i(k+1) = P(\theta_i \leq k+1|\mathbf{R}(k+1)) = \frac{1}{P(\mathbf{R}(k+1))} \cdot P(\mathbf{r}(k+1)|\theta_i \leq k+1) \cdot P(\mathbf{R}(k)|\theta_i \leq k+1) \cdot P(\theta_i \leq k+1). \quad (23)$$

Substituting from (20) to (22) into (23), we have

$$F_i(k+1) = \frac{1}{P(\mathbf{r}(k+1)|\mathbf{R}(k)) \cdot P(\mathbf{R}(k))} \cdot f_i(\mathbf{r}(k+1)) \cdot d\mathbf{r}(k+1) \cdot \frac{P(\theta_i \leq k+1|\mathbf{R}(k)) \cdot P(\mathbf{R}(k))}{P(\theta_i \leq k+1)} \cdot P(\theta_i \leq k+1) \quad (24)$$

$$= \frac{1}{P(\mathbf{r}(k+1)|\mathbf{R}(k))} \cdot f_i(\mathbf{r}(k+1)) \cdot d\mathbf{r}(k+1) \cdot P(\theta_i \leq k+1|\mathbf{R}(k)) \quad (25)$$

$$= \frac{f_i(\mathbf{r}(k+1)) \cdot dy(k+1) \cdot \phi_i(k)}{P(y(k+1)|\mathbf{R}(k))}. \quad (26)$$

Now,

$$P(\mathbf{r}(k+1)|\mathbf{R}(k)) = \sum_{j=0}^{m} P(\mathbf{r}(k+1)|\theta_j \leq k+1) \cdot P(\theta_j \leq k+1|\mathbf{R}(k)) \quad (27)$$

$$= \sum_{j=0}^{m} f_j(\mathbf{r}(k+1)) \cdot d\mathbf{r}(k+1) \cdot \phi_j(k), \quad (28)$$

so we can substitute into (26) to get

$$F_i(k+1) = \frac{f_i(\mathbf{r}(k+1)) \cdot \phi_i(k)}{\sum_{j=0}^{m} f_j(\mathbf{r}(k+1)) \cdot \phi_j(k)}. \quad (29)$$

Our induction is thus complete. □

2.2 RELATION TO A MULTIPLE HYPOTHESIS WALD SEQUENTIAL PROBABILITY RATIO TEST

If we restrict ourselves to the case where one hypothesis is correct for all time (*i.e.* we will never jump from one hypothesis to another), we reduce to the Wald [12] sequential probability ratio test:

$$F_i(k+1) = \frac{F_i(k) \cdot f_i(\mathbf{r}(k+1))}{\sum_{j=0}^{m} F_j(k) \cdot f_j(\mathbf{r}(k+1))}. \quad (30)$$

$$\phi_i(k) = F_i(k) + \tilde{p}_i \cdot (1 - F_i(k)), \qquad i \neq 0$$

$$\phi_0(k) = \prod_{i=1}^{m} \{1 - \phi_i(k)\}$$

$$F_i(k+1) = \frac{\phi_i(k) \cdot f_i(\mathbf{r}(k+1))}{\sum_{j=0}^{m} \phi_j(k) \cdot f_j(\mathbf{r}(k+1))}$$

Table 1: Summary of Shiryayev sequential probability ratio test

This is quite easy to show. Because there are no hypothesis jumps, $\tilde{p}_i = 0$ for all i, so

$$\phi_i(k) = F_i(k) + \tilde{p}_i \cdot (1 - F_i(k)) = F_i(k) \quad (31)$$

for all $i > 0$. Also, set of events $\{\mathcal{E}_i(k)\}$ is now mutually exclusive since the entire measurement sequence corresponds to a single hypothesis. Hence

$$\phi_0(k) = 1 - P\Big(\bigcup_{i=1}^{m} \mathcal{E}_i(k)\Big|\mathbf{R}(k)\Big) \quad (32)$$

$$= 1 - \sum_{i=1}^{m} P(\mathcal{E}_i(k)|\mathbf{R}(k)) \quad (33)$$

$$= 1 - \sum_{i=1}^{m} \phi_i(k) \quad (34)$$

$$= 1 - \sum_{i=1}^{m} F_i(k) \quad (35)$$

$$= F_0(k). \quad (36)$$

Thus, the recursive expression from the previous subsection becomes

$$F_i(k+1) = \frac{\phi_i(k) \cdot f_i(\mathbf{r}(k+1))}{\sum_{j=0}^{m} \phi_j(k) \cdot f_j(\mathbf{r}(k+1))} \quad (37)$$

$$= \frac{F_i(k) \cdot f_i(\mathbf{r}(k+1))}{\sum_{j=0}^{m} F_j(k) \cdot f_j(\mathbf{r}(k+1))}. \quad (38)$$

3 AN IMPROVED METHOD FOR INTEGER AMBIGUITY RESOLUTION

In this section we propose applying the statistical tests of the last section to an enlarged residual that uses both carrier phase and code information.

The method we propose may be used on either single or double differenced GPS pseudoranges. For simplicity, we will derive the method on double differenced data. The conversion of the method to single differenced data is straightforward.

Let us begin with the linearized carrier phase and code

measurement equations:

$$\nabla\Delta\tilde{\phi}(k)\lambda = \nabla\mathbf{H}(k)\delta\mathbf{x}(k) - \lambda\nabla\Delta\mathbf{N} + \nabla\eta_{car.}(k), \tag{39}$$

$$\nabla\delta\tilde{\rho}(k) = \nabla\mathbf{H}(k)\delta\mathbf{x}(k) + \nabla\eta_{code}(k), \tag{40}$$

where $\nabla\eta_{car.}$ and $\nabla\eta_{code}$ are independent zero-mean Gaussian random sequences with variances $\nabla\mathbf{V}_{car.}$ and $\nabla\mathbf{V}_{code}$, respectively. We can eliminate the terms dependent on $\delta\mathbf{x}$ by subtracting the code measurement from the carrier phase measurement, yielding the following relation:

$$\mathbf{r}^1(k) \triangleq \nabla\Delta\tilde{\phi}(k)\lambda - \nabla\delta\tilde{\rho}(k) = \nabla\eta_{car.}(k) - \nabla\eta_{code}(k) - \lambda\nabla\Delta\mathbf{N}. \tag{41}$$

Note that $\mathbf{r}^1$ is an independent Gaussian random sequence with mean $-\lambda\nabla\Delta\mathbf{N}$ and variance $(\nabla\mathbf{V}_{car.} + \nabla\mathbf{V}_{code})$.

Following the methodology of Park, *et al.* [6, 7], we can find an $\mathbf{E}(k)$ that is a left annihilator of $\nabla\mathbf{H}(k)$. Multiplying the carrier phase measurement on the left by $\mathbf{E}(k)$, we arrive at

$$\mathbf{r}^2(k) \triangleq \mathbf{E}(k)\nabla\Delta\tilde{\phi}(k)\lambda = \mathbf{E}(k)\nabla\eta_{car.}(k) - \lambda\mathbf{E}(k)\nabla\Delta\mathbf{N}. \tag{42}$$

Then $\mathbf{r}^2(k)$ is an independent Gaussian random sequence with mean $-\lambda\mathbf{E}(k)\nabla\Delta\mathbf{N}$ and variance $\mathbf{E}(k)\nabla\mathbf{V}_{car.}(k)\mathbf{E}^T(k)$.

Construct the vector $\mathbf{r}(k)$ as follows:

$$\mathbf{r}(k) = \begin{bmatrix}\mathbf{r}^1(k)\\ \mathbf{r}^2(k)\end{bmatrix} = \begin{bmatrix}\nabla\Delta\tilde{\phi}(k)\lambda - \nabla\delta\tilde{\rho}(k)\\ \mathbf{E}(k)\nabla\Delta\tilde{\phi}(k)\lambda\end{bmatrix} = \begin{bmatrix}\nabla\eta_{car.}(k) - \nabla\eta_{code}(k) - \lambda\nabla\Delta\mathbf{N}\\ \mathbf{E}(k)\nabla\eta_{car.}(k) - \lambda\mathbf{E}(k)\nabla\Delta\mathbf{N}\end{bmatrix}. \tag{43}$$

Then $\mathbf{r}(k)$ is an independent Gaussian random sequence with mean $\mathbf{m}_r(\nabla\Delta\mathbf{N}, k)$ and variance $\mathbf{V}_r(k)$ given by

$$\mathbf{m}_r(\nabla\Delta\mathbf{N}, k) = \begin{bmatrix}-\lambda\nabla\Delta\mathbf{N}\\ -\lambda\mathbf{E}(k)\nabla\Delta\mathbf{N}\end{bmatrix}, \tag{44}$$

$$\mathbf{V}_r(k) = \begin{bmatrix}\nabla\mathbf{V}_{code}(k) + \nabla\mathbf{V}_{car.}(k) & \cdots\\ \mathbf{E}(k)\nabla\mathbf{V}_{car.}(k) & \\ & \nabla\mathbf{V}_{car.}(k)\mathbf{E}^T(k)\\ & \mathbf{E}(k)\nabla\mathbf{V}_{car.}(k)\mathbf{E}^T(k)\end{bmatrix}. \tag{45}$$

Our proposed algorithm for integer ambiguity resolution is simply to apply MHWSPRT or MHSSPRT to $\mathbf{r}(k)$, with the hypothesis set $\{\mathcal{H}_1, \mathcal{H}_2, \ldots \mathcal{H}_m\}$ containing all of the values of $\nabla\Delta\mathbf{N}$ that are under consideration. We outline it in detail below.

Algorithm 3.1.

1. *Determine the values of* $\{\nabla\Delta\mathbf{N}_i\}$, $i = 1, 2, \ldots, m$ *under consideration as hypotheses. This can be done either by taking a set number of integers away from the code position estimate for each satellite, or by dividing the satellites into an independent and dependent set as in Park* et al.

2. *Initialize the probabilities* $F_i(0)$ *to their* a priori *values (For MHWSPRT, usually* $1/m$*, where* m *is the number of hypotheses under consideration. For MHSSPRT, usually* 1 *for the base hypothesis and* 0 *for the other hypotheses). Set k=0.*

3. *Take the* $(k+1)$*th measurements* $\nabla\Delta\phi(k+1)$ *and* $\nabla\delta\rho(k+1)$.

4. *Evaluate* $f_i(\mathbf{r}(k+1))$ *for* $i = 1, 2, \ldots, m$ *as follows:*

$$f_i(\mathbf{r}(k+1)) = \exp\{\mathbf{r}_i(k+1)^T\mathbf{V}_r(k+1)\mathbf{r}_i(k+1)\},$$

where

$$\mathbf{r}_i(k+1) = \begin{bmatrix}\mathbf{r}_i^1(k+1)\\ \mathbf{r}_i^2(k+1)\end{bmatrix},$$

and

$$\mathbf{r}_i^1(k+1) = \nabla\Delta\phi(k+1)\lambda - \nabla\delta\rho(k+1) + \lambda\nabla\Delta\mathbf{N}_i,$$

$$\mathbf{r}_i^2(k+1) = \mathbf{E}(k+1)\nabla\Delta\phi(k+1)\lambda + \lambda\mathbf{E}(k+1)\nabla\Delta\mathbf{N}_i(k+1).$$

Note that since all the hypotheses under consideration have identical covariances, the constant term preceding the exponent has been eliminated in the above expression.

5. *Calculate* $\{F_i(k+1)\}$ *using* $\{F_i(k)\}$ *and* $\{f_i(\mathbf{r}(k+1)\}$ *with either MHWSPRT or MHSSPRT, depending on whether we are determining the initial ambiguity or monitoring for cycle slips.*

6. *If we reach a desired threshold with any of the* $\{F_i(k+1)\}$*, declare the initial integer ambiguity and begin monitoring for cycle slips (MHWSPRT) or declare a cycle slip and reset the base hypothesis (MHSSPRT).*

7. *Go to step 3.*

For pedagogical reasons, we have used conventional L1 code pseudoranges in constructing the residual $\mathbf{r}^1(k)$. It is better in practice to use *narrowlane* code pseudorange combinations instead, because the combination of widelane carrier pseudoranges and narrowlane code pseudoranges yields a residual that contains no errors from ionospheric delay [13].

4 RESOLVING INTEGER AMBIGUITIES SEPARATELY

A key problem with the integer ambiguity resolution scheme we have presented, as well as with algorithms of the type proposed by Park *et al.* , is that the number of hypotheses that must be considered is large, as a result of the combinatorial relationship between the number of satellites and the number of integers to be examined per satellite. We propose a technique to alleviate this problem below.

Our approach attempts to construct residuals such that each residual is only affected by the integer ambiguity of a single satellite. Then the integer ambiguities of the satellites may be determined in parallel, with each parallel element making a choice from a small number of possible hypotheses.

Consider the carrier phase and code measurement equations again:

$$\begin{bmatrix} \nabla\Delta\tilde{\phi}(k)\lambda \\ \nabla\delta\tilde{\rho}(k) \end{bmatrix} = \begin{bmatrix} \nabla\mathbf{H}(k) & -\lambda\mathbf{I} \\ \nabla\mathbf{H}(k) & 0 \end{bmatrix} \begin{bmatrix} \nabla\delta\mathbf{x}(k) \\ \nabla\Delta\mathbf{N} \end{bmatrix} + \begin{bmatrix} \nabla\eta_{car.}(k) \\ \nabla\eta_{code}(k) \end{bmatrix}. \quad (46)$$

Suppose that we are only interested in the jth integer ambiguity $\nabla\Delta N^{(j)}$. If we exclude measurement equations with any of the other integer ambiguities, the equation above reduces to

$$\begin{bmatrix} \nabla\Delta\tilde{\phi}^{(j)}(k)\lambda \\ \nabla\delta\tilde{\rho}(k) \end{bmatrix} = \begin{bmatrix} \nabla\mathbf{h}^{(j)}(k) \\ \nabla\mathbf{H}(k) \end{bmatrix} \nabla\delta\mathbf{x}(k) - \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix} \nabla\Delta N^{(j)}\lambda + \begin{bmatrix} \nabla\eta^{(j)}_{car.}(k) \\ \nabla\eta_{code}(k) \end{bmatrix}. \quad (47)$$

Denote by $\mathbf{E}^{(j)}(k)$ the left annihilator of the matrix

$$\begin{bmatrix} \nabla\mathbf{h}^{(j)}(k) \\ \nabla\mathbf{H}(k) \end{bmatrix}.$$

Multiplying (47) by $\mathbf{E}^{(j)}(k)$ on the left yields

$$\mathbf{r}^{(j)}(k) \triangleq \mathbf{E}^{(j)}(k) \begin{bmatrix} \nabla\Delta\tilde{\phi}^{(j)}(k)\lambda \\ \nabla\delta\tilde{\rho}(k) \end{bmatrix} = \mathbf{E}^{(j)}(k) \begin{bmatrix} \nabla\eta^{(j)}_{car.}(k) \\ \nabla\eta_{code}(k) \end{bmatrix} - \mathbf{E}^{(j)}(k) \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix} \nabla\Delta N^{(j)}\lambda. \quad (48)$$

Hence, $\mathbf{r}^{(j)}(k)$ is a noise process, with distribution determined by the integer ambiguity $\nabla\Delta N^{(j)}$ and the joint distribution of $\eta_{car.}$ and η_{code}. The residual $\mathbf{r}^{(j)}(k)$ can thus be tested with a MHSPRT to determine which hypothesis is the correct value for $\nabla\Delta N^{(j)}$, independent of the other integer ambiguities.

5 A RESIDUAL FOR RESOLVING THE INTEGERS OF NEWLY ACQUIRED SATELLITES

If the integers corresponding to a number of satellites have been resolved and a new satellite comes into view, resolving the integer ambiguity of the new satellite is especially easy. Let $\nabla\Delta\mathbf{N} \in \mathbb{R}^{(M-1)}$ be the vector of resolved integer ambiguities corresponding to the carrier phase measurement vector $\nabla\Delta\phi(k) \in \mathbb{R}^{(M-1)}$. The carrier phase measurement corresponding to the new satellite is $\nabla\Delta\phi^{(M)}(k)$, and the integer ambiguity we seek to resolve is $\nabla\Delta N^{(M)}$. The measurement equations can then be written as

$$\begin{bmatrix} \nabla\Delta\phi^{(M)}(k) \\ (\nabla\Delta\phi(k) + \nabla\Delta\mathbf{N})\lambda \end{bmatrix} = \begin{bmatrix} \mathbf{h}^{(M)}(k) \\ \mathbf{H}(k) \end{bmatrix} \nabla\delta\mathbf{x}(k) + \begin{bmatrix} -1 \\ \mathbf{0} \end{bmatrix} \nabla\Delta N^{(M)} + \begin{bmatrix} \eta^{(M)}_{car.} \\ \eta_{car.} \end{bmatrix}. \quad (49)$$

Construct a measurement residual $\mathbf{r}^{(M)}(k)$ by multiplying (49) by $\mathbf{E}^{(M)}(k)$, the left annihilator of the matrix

$$\begin{bmatrix} \mathbf{h}^{(M)}(k) \\ \mathbf{H}(k) \end{bmatrix}:$$

$$\mathbf{r}^{(M)}(k) \triangleq \mathbf{E}^{(M)}(k) \begin{bmatrix} \nabla\Delta\phi^{(M)}(k) \\ (\nabla\Delta\phi(k) + \nabla\Delta\mathbf{N})\lambda \end{bmatrix} = \mathbf{E}^{(M)}(k) \begin{bmatrix} \eta^{(M)}_{car.} \\ \eta_{car.} \end{bmatrix} + \mathbf{E}^{(M)}(k) \begin{bmatrix} -1 \\ \mathbf{0} \end{bmatrix} \nabla\Delta N^{(M)}. \quad (50)$$

The measurement residual $\mathbf{r}^{(M)}(k)$ is a random noise sequence, with distribution determined by $\nabla\Delta N^{(M)}$ and the joint distribution of $\eta^{(M)}_{car.}$ and $\eta_{car.}$. Taking advantage of the small number of possible hypotheses under consideration and the low noise associated with $\mathbf{r}^{(M)}(k)$, a MHWSPRT can quickly determine the correct value of the new integer ambiguity $\nabla\Delta N^{(M)}$.

6 EXPERIMENTS

In this section, we evaluate the performance of the methods derived in the previous sections.

6.1 STATIONARY SINGLE ANTENNA EXPERIMENT

We first constructed an experimental apparatus in which the integer ambiguity was known. We connected two Ashtech model Z-12 GPS receivers to a single Sensor Systems model S67-1575-96 L1/L2 active antenna, so that the integer bias was known to be zero for every carrier phase measurement. We then compared the results of a Wald test using the residual (43) to those using the carrier-only residual (42).

The data sequences we measured contained observations of at least seven satellites. To test the algorithm, we eliminated some of the measurements, so that there were either five or six visible satellites visible. These reduced data sets were double differenced and widelaned, and then processed by the integer ambiguity resolution algorithm. With either residual, the algorithm always correctly concluded that the integer biases were zero. We have plotted the time history of the maximum valued F_i's for both the five and six satellite cases in Figures 1 and 2. Note that while the addition of GPS code measurements only slightly improves the convergence in the six satellite case, the improvement in the five satellite case is quite significant.

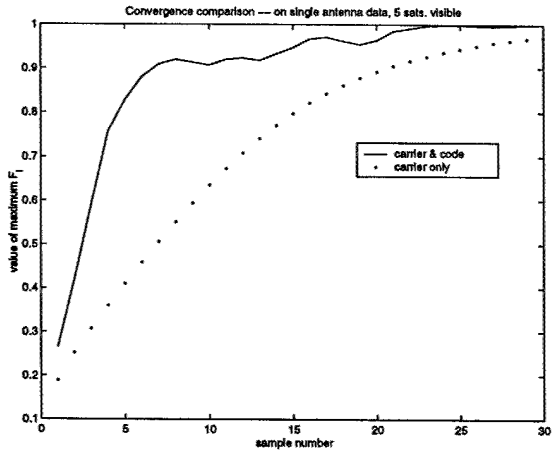


Figure 1: Comparison between different residuals: Maximum value of F_i vs. time, 5 satellites visible

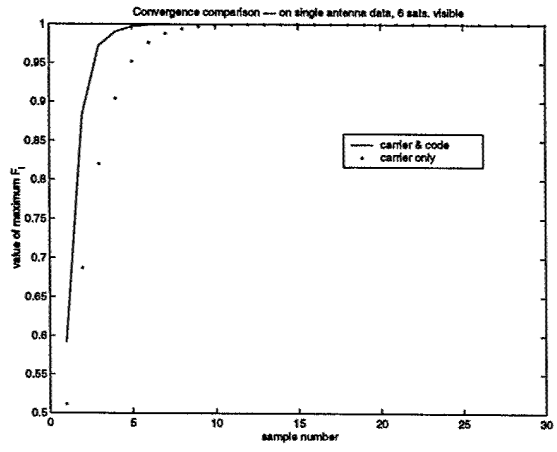


Figure 2: Comparison between different residuals: Maximum value of F_i vs. time, 6 satellites visible

Finally, we looked at the performance of the Wald test applied to the error residual in (41). Using the same original measurement sequence, we reduced the number of visible satellites in the data to four before processing them. In this case, the correct hypothesis was not determined until after 75 measurements, due to a large excursion that one of the code measurements took from its mean. In Figure 3 we plotted the time history of the most probable hypothesis (the correct hypothesis number in this case was 14). In Figure 4 we plotted the time history of the maximum valued F_i.

Note that when our algorithm assumed that the standard deviation of the code measurements was 1 meter, the Wald test told us that it was 100% sure that an incorrect hypothesis was correct! To avoid this overconfidence problem in our algorithm, we increased the standard deviation of the code measurements to 2 meters. While this did not decrease the time at which the correct hypothesis was declared the most probable, it did avoid the problem of declaring the wrong hypothesis correct.

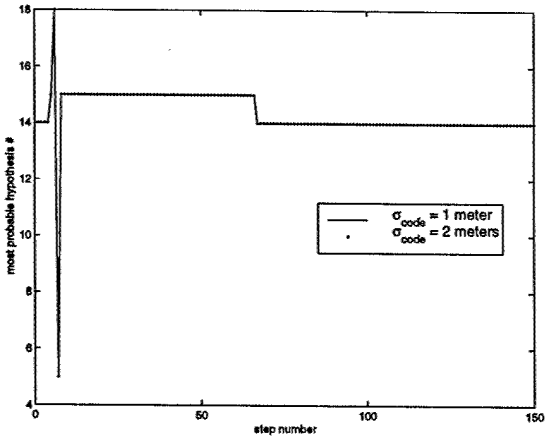


Figure 3: Most probable hypothesis vs. time, 4 satellites visible. The correct hypothesis number is 14

6.2 DYNAMIC TESTS

In this subsection, several different techniques for resolving integer ambiguity were applied to the same data set. The GPS data was collected using a test rig in which two Sensor Systems model S67-1575-96 L1/L2 active GPS antennae were placed a fixed distance from each other (2.3 meters). The test rig was mounted to a car, which was driven at moderate speed for several minutes while the GPS measurements were recorded from Ashtech model Z-12 GPS receivers. The results of the integer ambiguity resolution schemes could be readily verified, as filtering of the code measurements using the resolved integers generated a distance estimate between the two antennae, which would not match the true distance unless

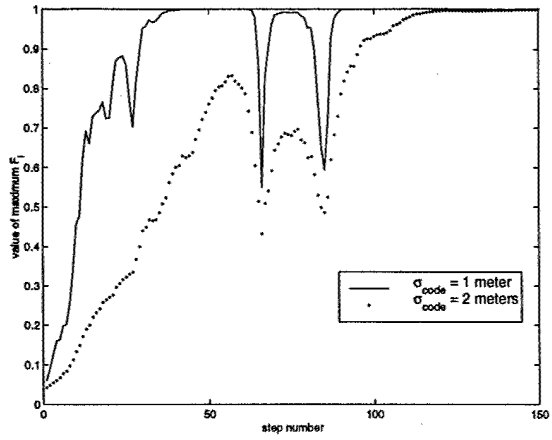


Figure 4: Maximum value of F_i vs. time, 4 satellites visible

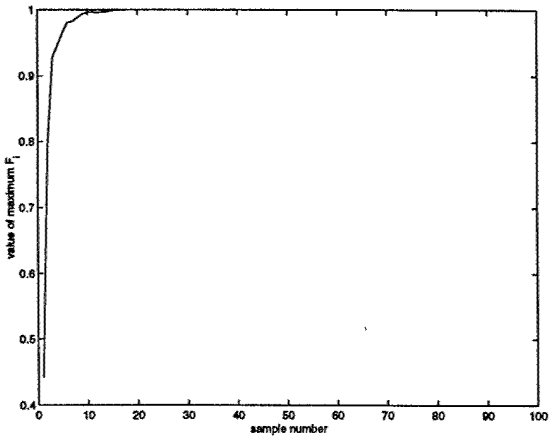


Figure 5: Maximum value of F_i vs. time, residual uses 5 double differenced carrier and 5 double differenced code measurements

the resolved integers were correct. The experimental data was collected beginning Friday, October 13, 2000 at 19:22:33.5 PST.

6.2.1 WALD TEST USING MULTIPLE CARRIER MEASUREMENTS AND MULTIPLE CODE MEASUREMENTS

A Wald test using a residual generated by 5 double differenced widelane carrier measurements and 5 double differenced narrowlane code measurements quickly resolved the correct integer ambiguities. There were five integer values under consideration for each double differenced carrier measurement, for a total of $5^5 = 3125$ different hypotheses. The noise in each of the single differenced widelane carrier phase measurements was assumed to be a zero mean Gaussian white noise process with a $2\sqrt{2}$ cm. standard deviation. The noise in each single differenced widelane code measurement was assumed to be a zero mean Gaussian white noise process with a $\sqrt{2}$ m. standard deviation. Figure 5 shows the time history of the probability of the most probable hypothesis.

6.2.2 WALD TEST USING A SINGLE CARRIER MEASUREMENT AND MULTIPLE CODE MEASUREMENTS

When a Wald test was performed using the residual (48), the correct integer ambiguities were again resolved, albeit more slowly. The residual for each integer ambiguity used a single double differenced widelane carrier phase measurement and as many double differenced narrowlane code measurements as were available at each epoch (between 5 and 7). There were 9 integer values under consideration for each satellite. The noise in the double differenced carrier measurement was assumed to be a zero mean Gaussian white noise process with standard deviation $2\sqrt{2}$ cm. The noise in each double differenced code measurement was assumed to be a zero mean Gaussian white noise process with standard deviation 2 m. Figure 6 shows the time history of the probability of the most probable hypothesis for one integer ambiguity, a pattern that was typical for all the ambiguities for which we searched. The price of the convenience, simplicity and low computational cost associated with evaluating the ambiguities separately was slower convergence of the Wald test.

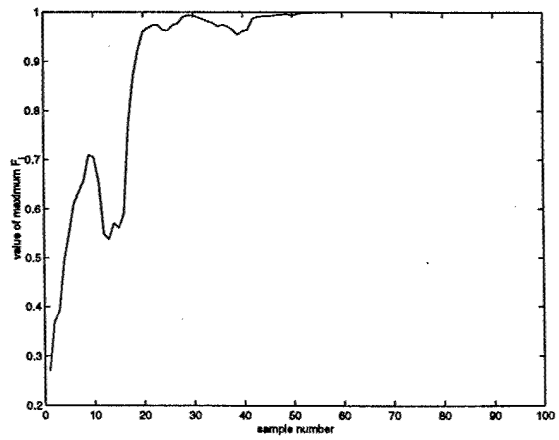


Figure 6: Maximum value of F_i vs. time, evaluating integer ambiguities separately

6.2.3 SHIRYAYEV TEST USING MULTIPLE CARRIER MEASUREMENTS AND MULTIPLE CODE MEASUREMENTS

Since there were no cycle slips in the observed data, we artificially introduced one into the measurement sequence in order to test our cycle slip monitoring scheme. We injected the cycle slip into one of the carrier phase measurements at the 100th epoch. We then ran a Shiryayev test on the same residual as before, using the integers resolved by the Wald test as the nominal hypothesis. The assumed standard deviation of the carrier measurements was increased to 8 cm., as lower values made the test too eager to declare a cycle slip. Figure 7 shows the time history of the most probable hypothesis. The correct hypothesis number is 313 before 100 samples, and it is 312 after 100 samples. Figure 8 plots the time history of the probability of the most probable hypothesis. The Shiryayev test detected that a cycle slip had occurred immediately, but it took 30 samples until it identified the correct hypothesis as the most probable one, and an additional 20 samples until the probability of that hypothesis was near 100%.

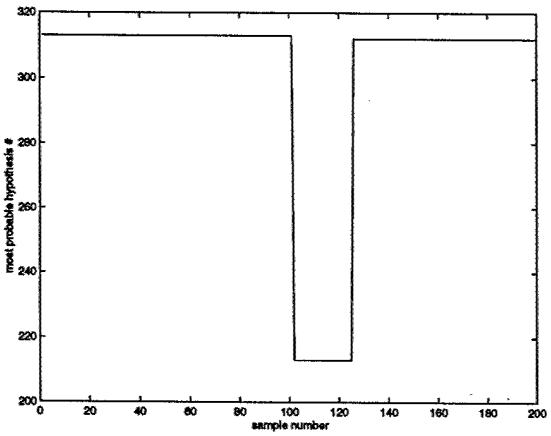


Figure 7: Most probable hypothesis vs. time, correct value is 313 before 100 samples and 312 afterwards

7 CONCLUSION

This paper outlines the integer ambiguity problem for GPS and describes some new methods for resolving the integer ambiguities and for detecting cycle slips. The main contribution is the application of the Wald and Shiryayev multiple hypothesis sequential probability ratio tests (MHSPRTs) to dynamically compute the conditional probabilities of members of a set of integer hypotheses being correct. Since the MHSPRTs require evaluations of the probability density functions of the measurement residuals, the expressions we constructed for the probability density functions of several residuals are also of interest. A large

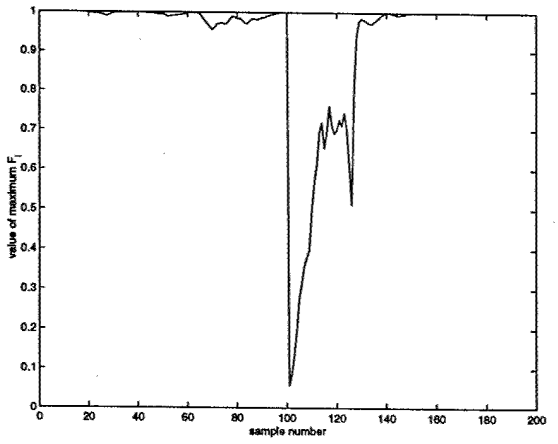


Figure 8: Maximum value of F_i vs. time, cycle slip occurs after 100 samples

number of experiments, performed on both simulations and actual GPS measurements, demonstrates the effectiveness of our methods.

Although there are many other methods for identifying the initial integer biases and for detecting cycle slips, the methods presented here possess several advantages over their peers. Chief among these advantages is the presentation of information about the integer hypotheses in a probabilistic framework, instead of the cumulative sum approach used by competing methods. This probabilistic framework allows for easy accommodation of events such as the introduction of new hypotheses. Other advantages of our methods are cycle slip detection that automatically determines the new integer hypothesis after the slip (rather than simply announcing that a cycle slip has occurred), provisions for the easy accommodation of non-Gaussian measurement noises, and efficient computation due to the recursive nature of the MHSPRTs.

The most significant disadvantage of our method is that only a finite number of hypotheses may be considered. Further, the computational cost increases as the number of hypotheses increases. We have presented a residual that drastically limits the number of hypotheses under consideration, but the price is an increase in the noise of the residual. This provides a GPS receiver designer with the option of trading between the computational load for each epoch versus the number of epochs of data required to accurately determine the integer ambiguities.

In closing, note that our methodology need not be exclusive of other techniques for integer ambiguity resolution. For instance, a small set of admissible integer ambiguity hypotheses can obtained via Teunissen's LAMBDA method [1, 2]. These hypotheses could then be analyzed using our techniques. In making this set small, the LAMBDA method greatly reduces the computational time required by our techniques.

ACKNOWLEDGMENTS

Our research was supported by the United States Air Force Office of Scientific Research under Grant F496200-00-0154, and by NASA Dryden Flight Research Center under Grant NAS4-00051.

REFERENCES

[1] P. J. G. Teunissen, "A new method for fast carrier phase ambiguity estimation," in *IEEE 1994 Position Location and Navigation Symposium*, (Las Vegas, NV), pp. 562–73, IEEE, April 1994.

[2] P. J. G. Teunissen, "The least-squares ambiguity decorrelation adjustment: a method for fast GPS integer ambiguity estimation," *Journal of Geodesy*, vol. 70, pp. 65–82, November 1995.

[3] C. C. Counselman and S. A. Gourevitch, "Miniature interferometer terminals for earth surveying: Ambiguity and multipath with global positioning system," *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-19, pp. 244–52, October 1981.

[4] S. J. Corbett and P. A. Cross, "GPS single epoch resolution," *Survey Review*, vol. 33, pp. 149–60, July 1995.

[5] Y. M. Al-Haifi, S. J. Corbett, and P. A. Cross, "Performance evaluation of GPS single-epoch on-the-fly ambiguity resolution," *Navigation. Journal of the Institute of Navigation*, vol. 44, pp. 479–87, Winter 1997-1998.

[6] C. Park, I. Kim, J. G. Lee, and G.-I. Lee, "Efficient technique to fix GPS carrier phase integer ambiguity on-the-fly," *IEE Proceedings - Radar, Sonar and Navigation*, vol. 144, pp. 148–55, June 1997.

[7] C. Park, I. Kim, J. G. Lee, and G.-I. Lee, "Efficient ambiguity resolution using constraint equation," in *IEEE 1996 Position Location and Navigation Symposium – PLANS '96*, (Atlanta, GA), pp. 277–84, IEEE, IEEE, April 1996.

[8] S. P. Mertikas and C. Rizos, "On-line detection of abrupt changes in the carrier-phase measurements of GPS," *Journal of Geodesy*, vol. 71, pp. 469–82, July 1997.

[9] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100–115, 1954.

[10] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Statistics*, vol. 42, pp. 1897–1908, 1971.

[11] D. P. Malladi and J. L. Speyer, "A generalized Shiryayev sequential probability ratio test for change detection and isolation," *IEEE Transactions on Automatic Control*, vol. 44, pp. 1522–34, August 1999.

[12] A. Wald, *Sequential Analysis*. Wiley mathematical statistics series, New York: J. Wiley and sons, Inc., 1947.

[13] G. Blewitt, "An automatic editing algorithm for GPS data," *Geophysical Research Letters*, vol. 17, pp. 199–202, March 1990.